# The Standardization of the Digital Dialect: Artificial Intelligence, Epistemic Injustice, and the Future of Linguistic Diversity

**Elizabeth Njeri Ngigi**

**Orangeburg Wilkinson High School, Orangeburg, South Carolina, United States of America**

**DOI: https://doi.org/10.51584/IJRIAS.2026.11050080**

**Received: 10 May 2026; Accepted: 16 May 2026; Published: 01 June 2026**

## ABSTRACT

Language has historically functioned as both a communicative system and a repository of cultural identity, social memory, and collective belonging. In the contemporary digital era, Artificial Intelligence (AI) technologies increasingly mediate communication practices through predictive text systems, speech recognition software, algorithmic moderation, and Large Language Models (LLMs). Although such technologies have improved efficiency, accessibility, and multilingual interaction, emerging scholarship suggests that AI systems may also reproduce existing linguistic hierarchies by privileging dominant and standardized language forms while marginalizing dialectal and non-standard varieties (Helm et al., 2024; Hofmann et al., 2024).

This theoretical paper examines how AI-driven communication systems may contribute to processes of linguistic standardization and dialect marginalization within digital environments. Drawing upon Fricker's (2007) theory of epistemic injustice, the paper introduces the concept of the Digital Language Divide to explain disparities in computational recognition, linguistic legitimacy, and technological visibility among speech communities. Unlike deterministic arguments that portray technology solely as a force of linguistic erosion, this paper adopts a balanced perspective acknowledging that AI technologies may simultaneously expand communication access while also reinforcing structural inequalities in language representation.

Recent empirical studies demonstrate that some AI systems exhibit dialect prejudice, particularly toward speakers of African American English and other non-standard language varieties (Hofmann et al., 2024). Additionally, emerging scholarship on techno-linguistic bias argues that language technologies frequently prioritize dominant linguistic norms embedded within training datasets (Helm et al., 2024). Building upon these findings, this paper theorizes how algorithmic standardization may gradually influence linguistic practices, educational expectations, and communicative participation within digitally mediated societies.

This paper seeks to establish a conceptual foundation for future interdisciplinary inquiry into the sociolinguistic consequences of AI-mediated communication. The study concludes by proposing directions for future empirical research, policy development, and inclusive AI design frameworks capable of supporting linguistic diversity within digital ecosystems.

**Keywords:** Artificial Intelligence, Digital Language Divide, Linguistic Diversity, Epistemic Injustice, Dialect Bias, Sociolinguistics, Language and Technology, Algorithmic Standardization

## INTRODUCTION

Language remains deeply intertwined with identity, social belonging, historical memory, and cultural continuity. Sociolinguistic scholarship has consistently demonstrated that dialects and non-standard linguistic varieties are not deficient forms of communication but rather legitimate systems shaped by community experience, history, and social context (Lippi-Green, 2012; Milroy & Milroy, 2012). Nevertheless, institutional

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue V May 2026

structures have historically privileged standardized language forms in educational, political, and professional spaces, often associating them with authority, intelligence, and legitimacy.

In recent years, Artificial Intelligence technologies have become increasingly influential in shaping everyday communication practices. AI-powered systems now mediate written communication, online interaction, translation, educational feedback, and speech recognition on an unprecedented scale. Tools such as ChatGPT, predictive text applications, grammar correction software, and automated transcription systems increasingly influence how individuals write, speak, and interact within digital environments.

Despite the benefits associated with these technologies, emerging research suggests that AI systems are not linguistically neutral. Because most Large Language Models are trained using datasets dominated by standardized forms of language, they may reproduce existing social and linguistic inequalities embedded within those datasets (Gallegos et al., 2024). Recent empirical studies have demonstrated that AI systems can display measurable forms of dialect bias and linguistic discrimination. Hofmann et al. (2024), for instance, found that language models produced significantly more negative judgments toward speakers of African American English than toward speakers of standardized English varieties. Similarly, Fleisig et al. (2024) reported that conversational AI systems frequently generated stereotypical, dismissive, or less comprehensible responses when interacting with non-standard dialects.

However, it is important to avoid overly deterministic interpretations of technological development. AI technologies have also created opportunities for multilingual communication, accessibility support, language preservation initiatives, and broader participation in digital knowledge economies. For example, AI-driven translation systems and speech technologies have improved access to communication for many marginalized linguistic communities. Consequently, the relationship between AI and language diversity should not be understood as uniformly harmful or inevitably destructive.

This paper therefore adopts a balanced theoretical position. Rather than arguing that AI technologies directly cause dialect extinction, the study examines how AI systems may contribute to subtle pressures toward linguistic conformity by privileging computationally recognizable forms of communication. Drawing upon Epistemic Injustice Theory (Fricker, 2007), the paper introduces the concept of the Digital Language Divide to explain how disparities in technological recognition may influence broader questions of linguistic legitimacy, visibility, and participation within digital society

**Understanding the Digital Language Divide**

Discussions surrounding the digital divide have traditionally focused on inequalities in access to technology, internet connectivity, and digital literacy. However, as AI increasingly mediates communication itself, another form of inequality may be emerging ; one centered not on access to technology, but on linguistic recognition within technological systems.

This paper introduces the term *Digital Language Divide* to describe disparities in the way languages and dialects are recognized, represented, and validated within AI-driven communication environments. In this context, some forms of language become more “visible” to technology than others.

Languages and dialects that are well represented in AI training datasets tend to function more smoothly within digital systems. Speech recognition tools understand them more accurately, predictive text systems respond to them more effectively, and grammar correction tools tend to validate them as linguistically acceptable. By contrast, dialects that receive less representation may encounter persistent misunderstanding, correction, mistranscription, or exclusion.

The issue is not simply technical. Language carries social meaning. When technological systems repeatedly fail to recognize certain dialects, the speakers of those dialects may begin to experience forms of linguistic invisibility within digital spaces. Over time, such patterns may reinforce the idea that some ways of speaking are more legitimate, intelligent, or socially valuable than others.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue V May 2026

This concern becomes especially significant because digital communication increasingly shapes educational participation, professional interaction, social networking, and public discourse. As AI systems become more deeply integrated into everyday life, technological recognition may gradually influence broader perceptions of linguistic legitimacy.

The Digital Language Divide therefore reflects more than computational limitations; it points to deeper questions about power, representation, and cultural inclusion in the digital age.

Standardizing digital dialects through Artificial Intelligence risks flattening the world's rich linguistic diversity into homogeneous, often Anglo-centric norms. This algorithmic homogenization drives epistemic injustice by obscuring marginalized worldviews, ultimately restricting the credibility and interpretive resources of non-dominant language communities.

**The Mechanics of Algorithmic Standardization**

The "Machine Translationese" Effect: AI models trained predominantly on heavily edited, English-language, or majoritarian corpora develop an algorithmic bias. This leads to the loss of peripheral vocabulary, complex pragmatics, and stylistic nuance.

The Digital Language Divide: Over 6,000 of the world's languages remain unsupported or "digitally disadvantaged". When AI translates or generates content in low-resource contexts, it frequently produces grave inaccuracies because it maps minority dialects onto dominant frameworks.

## AI AND THE PRESSURE TOWARD LINGUISTIC STANDARDIZATION

One of the defining characteristics of AI systems is their dependence on patterns and predictability. Large Language Models operate by identifying the most statistically probable forms of language within massive datasets. Consequently, the language forms most frequently represented within these datasets become normalized within AI outputs.

This process may unintentionally encourage linguistic standardization.The push toward a standardized digital dialect presents several core weaknesses, alongside profound ethical and cultural implications:

**The Mechanics of the Linguistic Monoculture**

Anglo-Centric Training Data: Large Language Models (LLMs) are predominantly trained on English and a handful of high-resource languages. This skews the foundational data and enforces Western conceptual frameworks as the "default" standard of logic and communication.

The "Great Linguistic Flattening": AI models struggle to capture the nuances of long-tail, under-resourced languages. As users increasingly rely on AI for translation, grammar, and content generation, these algorithms subtly pressure speakers to abandon their native phrasing in favor of homogenized, algorithmic outputs.

For example, predictive text systems often guide users toward standardized grammar and vocabulary choices. Autocorrection tools routinely replace dialectal spellings and expressions with standardized alternatives. In many cases, users unconsciously adapt their language to align with technological expectations because doing so improves communication efficiency within digital systems.

Speech recognition technologies also illustrate this issue. Numerous users from marginalized linguistic communities report that voice assistants and transcription systems struggle to understand accents, dialects, and culturally specific speech patterns. Such experiences may appear minor at first glance, but repeated technological misunderstanding can gradually shape how individuals perceive the value and acceptability of their own linguistic identities.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue V May 2026

In educational environments, AI-powered writing tools increasingly influence students' understanding of what constitutes "correct" language. As students rely more heavily on AI-driven grammar systems, there is a possibility that dialectal expression may become further stigmatized within academic spaces.

Similarly, professional communication platforms may reward standardized linguistic performance while subtly penalizing dialectal variation. In this sense, AI systems may contribute to environments where linguistic conformity becomes socially advantageous.

Importantly, this paper does not suggest that AI developers intentionally design systems to suppress dialects. Rather, the argument is that computational efficiency and linguistic diversity may sometimes exist in tension. AI systems function most effectively when language follows predictable patterns, yet human language is inherently diverse, fluid, and culturally embedded.

This tension raises important questions about the future relationship between technology and linguistic identity.

## EPISTEMIC INJUSTICE AND DIGITAL SILENCING

To better understand the broader implications of dialect marginalization within AI systems, this paper draws upon the theory of epistemic injustice developed by Miranda Fricker. Epistemic injustice refers to forms of unfairness that individuals experience specifically in their role as knowers, communicators, and participants in social understanding.

The theory becomes particularly relevant when examining how AI systems engage with language.

### Testimonial Injustice

Testimonial injustice occurs when prejudice causes a speaker to be viewed as less credible or less trustworthy. Historically, speakers of certain dialects have often faced assumptions that they are less educated, less intelligent, or less professional than speakers of standardized language varieties.

AI systems may unintentionally reproduce such biases if they consistently struggle to interpret or validate certain dialects. When a technological system repeatedly misrecognizes a person's speech or flags their language as incorrect, it may subtly reinforce harmful social stereotypes surrounding linguistic legitimacy.

Over time, speakers may begin to feel pressure to alter their speech in order to gain acceptance within digitally mediated spaces.

### Hermeneutical Injustice

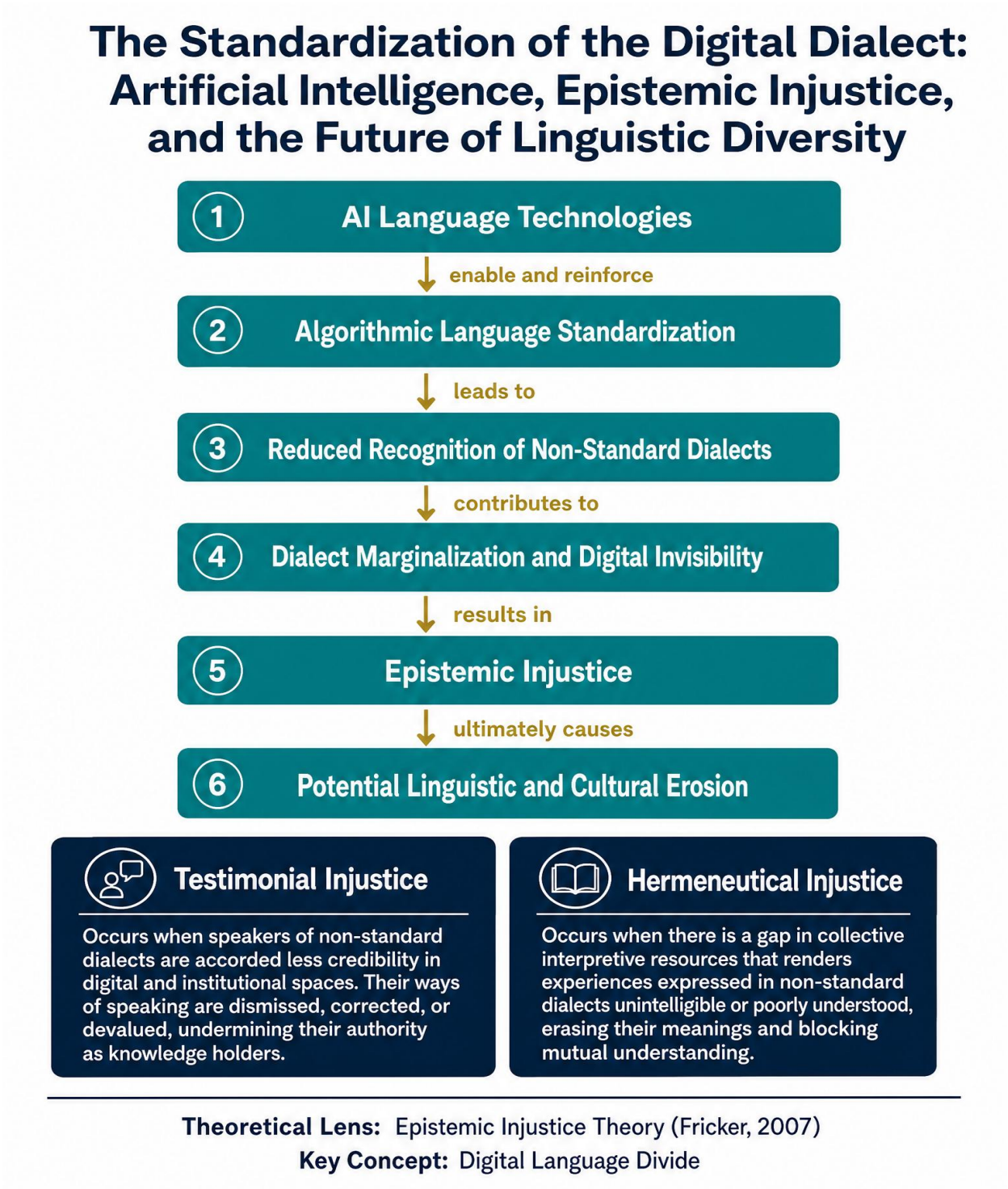
Hermeneutical injustice occurs when certain experiences or identities are inadequately represented within systems of understanding. Within AI systems, this may happen when dialects are underrepresented in training datasets, making them difficult for algorithms to interpret effectively.

When technological systems cannot fully recognize a linguistic community, those speakers risk becoming digitally invisible. Their forms of expression remain outside the dominant computational framework through which communication increasingly occurs.

This paper argues that such invisibility should not be viewed merely as a technical oversight. Instead, it raises broader concerns regarding whose voices are included, recognized, and valued within the future digital public sphere.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue V May 2026

## THEORETICAL FRAMEWORK




### Epistemic Injustice Theory

This paper is grounded in Epistemic Injustice Theory, which examines how systems of power influence whose knowledge, language, and experiences are recognized as legitimate within society.

The theory provides a useful framework for understanding how AI technologies may shape linguistic inclusion and exclusion. As AI systems increasingly mediate communication, education, and access to information, they begin to function not simply as technological tools but also as social gatekeepers.

Within this context, computational recognition may become closely connected to social recognition. Dialects that AI systems process effectively may gain increased legitimacy within digital environments, while those that remain poorly represented may experience exclusion or marginalization.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue V May 2026

The theory therefore allows this paper to explore the relationship between technological visibility and linguistic legitimacy. It also highlights the possibility that language-based inequalities within AI systems may reproduce broader social hierarchies already present within society.

Rather than viewing dialect bias solely as a technical flaw, Epistemic Injustice Theory encourages a deeper examination of how technology may shape participation, identity, and communicative power in the digital age.

The framework begins with AI language technologies, including Large Language Models, predictive text systems, grammar correction software, automated moderation systems, and speech recognition technologies. These systems are typically trained on datasets dominated by standardized and institutionally privileged forms of language.

The second stage of the framework involves algorithmic language standardization. Because AI systems operate through statistical pattern recognition, frequently occurring linguistic structures become prioritized within computational outputs. Consequently, dominant linguistic forms receive greater technological validation and smoother functional performance.

The third stage concerns reduced recognition of non-standard dialects. Empirical research increasingly demonstrates that AI systems may misunderstand, mis transcribe, or inadequately process dialectal speech patterns (Hofmann et al., 2024). Such limitations may affect communication quality, accessibility, and perceptions of linguistic legitimacy.

The fourth stage involves dialect marginalization and digital invisibility. When dialect speakers repeatedly encounter technological misunderstanding or correction, certain forms of speech may become less visible and less socially validated within digital environments.

The framework then incorporates Fricker's (2007) concepts of testimonial injustice and hermeneutical injustice. Testimonial injustice occurs when dialect speakers are perceived as less credible or less authoritative due to linguistic prejudice. Hermeneutical injustice emerges when dialectal experiences remain insufficiently represented within dominant technological and communicative systems.

## THE NEED FOR FUTURE RESEARCH

Although conversations surrounding AI ethics have expanded significantly in recent years, comparatively little attention has been directed toward the sociolinguistic implications of AI-driven communication systems. Existing research often focuses on issues such as misinformation, privacy, political bias, and racial discrimination, while questions concerning dialect diversity remain underexplored.

This paper therefore calls for further interdisciplinary research examining the relationship between AI systems and linguistic diversity.

Future studies could investigate whether prolonged interaction with AI technologies influences individuals to modify or abandon dialectal speech patterns in favor of standardized forms. Longitudinal studies may help determine whether younger generations increasingly adapt their language to align with AI-compatible communication styles.

Research is also needed within educational contexts. As AI-powered writing tools become more common in schools and universities, scholars should examine how these systems shape students' attitudes toward dialects and linguistic legitimacy.

Another important area concerns speech recognition technologies and automated moderation systems. Empirical studies could explore whether certain dialects experience disproportionately high error rates or negative algorithmic treatment within digital platforms.

Comparative international research may further investigate how AI systems affect minority languages and indigenous linguistic communities across different regions of the world. Such studies would contribute to broader discussions concerning digital inclusion, cultural preservation, and linguistic justice.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue V May 2026

# FUTURE RESEARCH DIRECTIONS AND IMPLEMENTATION STRATEGIES

Future research should move beyond theoretical concerns and empirically investigate how AI-mediated communication influences linguistic practices across different social and educational settings. Longitudinal studies may help determine whether younger generations increasingly adapt their language patterns to align with AI-compatible communication norms.

Research is also needed regarding educational technologies. As AI-powered writing tools become more integrated into schools and universities, scholars should examine how these systems shape student perceptions of linguistic correctness, dialect legitimacy, and academic identity.

International comparative studies would further strengthen understanding of how AI systems affect minority languages and indigenous linguistic communities across different sociocultural contexts. For example, several African, Asian, and Indigenous languages remain significantly underrepresented in major AI datasets, potentially contributing to broader forms of digital exclusion.

Policy-oriented research should additionally examine strategies for improving linguistic inclusivity within AI development. Possible implementation strategies include:

1. Expanding dialectal and multilingual representation within AI training datasets.
2. Developing fairness benchmarks specifically designed for dialect recognition.
3. Incorporating sociolinguistic expertise into AI ethics frameworks.
4. Supporting public funding for minority-language digital preservation initiatives.
5. Encouraging transparency regarding language representation within AI systems.

International organizations and policy frameworks may also provide useful models. UNESCO's growing emphasis on inclusive digital transformation and linguistic diversity highlights the importance of culturally responsive AI development within global technology governance.

At the same time, implementation challenges remain significant. Expanding dialect representation within AI systems requires extensive linguistic datasets, financial investment, ethical oversight, and culturally informed annotation practices. Additionally, balancing computational efficiency with linguistic diversity presents technical challenges for developers working with large-scale AI systems

Consequently, AI governance should be understood as part of a broader sociocultural effort to promote equitable communication and digital participation.

# CONCLUSION

The increasing presence of Artificial Intelligence within human communication raises important questions about the future of language diversity in digital society. While AI technologies provide significant opportunities for efficiency, accessibility, and global interaction, they may also contribute to subtle pressures toward linguistic conformity.

This paper has argued that AI systems may be shaping a new form of linguistic hierarchy in which computational recognition becomes increasingly tied to social legitimacy. Through the concept of the Digital Language Divide, the discussion highlights how dialects and marginalized speech varieties may experience forms of technological invisibility within AI-mediated environments.

Drawing upon Epistemic Injustice Theory, the paper further suggests that failures of technological recognition are not merely technical concerns. They are also social and cultural concerns connected to representation, belonging, and communicative power.

The purpose of this paper has not been to offer definitive empirical conclusions, but rather to open a theoretical conversation regarding the possible long-term implications of AI-driven linguistic standardization. As AI

technologies continue to shape communication practices around the world, questions concerning dialect recognition, linguistic diversity, and cultural preservation may become increasingly urgent.

Ultimately, the future of language in the digital age may depend not only on human communities preserving their linguistic identities, but also on whether technological systems are designed to recognize, respect, and accommodate the full diversity of human expression.

## REFERENCES

1. Abdalla, M., Wahle, J. P., Ruas, T., Névéol, A., & Ducel, F. (2023). The elephant in the room: Analyzing the presence of Big Tech in natural language processing research. Association for Computational Linguistics.
2. Buolamwini, J. (2023). Unmasking AI: My mission to protect what is human in a world of machines. Random House.
3. Fleisig, E., Smith, G., Bossi, M., Rustagi, I., Yin, X., & Klein, D. (2024). Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. arXiv. https://arxiv.org/abs/2406.08818
4. Fricker, M. (2007). Epistemic injustice: Power and the ethics of knowing. Oxford University Press.
5. Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., & others. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), 1097–1179.
6. Helm, P., Bella, G., Koch, G., & Giunchiglia, F. (2024). Diversity and language technology: How language modeling bias causes epistemic injustice. Ethics and Information Technology, 26(1), 1–15. https://doi.org/10.1007/s10676-023-09742-6
7. Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. Nature, 633, 147–154. https://doi.org/10.1038/s41586-024-07856-5
8. Kay, J., Kasirzadeh, A., & Mohamed, S. (2024). Epistemic injustice in generative AI. arXiv. https://arxiv.org/abs/2408.11441
9. Lawrence, H. (2022). Siri disciplines: AI assistants and the bias against accented speech. Journalism & Mass Communication Quarterly, 99(2), 456–478.
10. Lippi-Green, R. (2012). English with an accent: Language, ideology, and discrimination in the United States (2nd ed.). Routledge.
11. Medina, J. (2013). The epistemology of resistance: Gender and racial oppression, epistemic injustice, and resistant imaginations. Oxford University Press.
12. Milroy, J., & Milroy, L. (2012). Authority in language: Investigating standard English (4th ed.). Routledge.
13. Mollema, W. J. T. (2025). A taxonomy of epistemic injustice in the context of AI and the case for generative hermeneutical erasure. arXiv. https://arxiv.org/abs/2504.07531
14. Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. New York University Press.
15. Olaniyan, Y. D., Martins, M. O., & Al Maqrashi, R. H. (2026). Generative artificial intelligence and epistemic (in)justice: Perspectives from higher education students in the Global North and South. Frontiers in Human Dynamics, 8, Article 1790324.
16. Pretorius, L., Huynh, H.-H., Pudyanti, A. A. A. R., Li, Z., & Noori, A. Q. (2025). Empowering international PhD students: Generative AI, Ubuntu, and the decolonisation of academic communication. The Internet and Higher Education, 64, 100942.
17. UNESCO. (2023). Guidance for generative AI in education and research. UNESCO Publishing.
18. Van Dijck, J. (2013). The culture of connectivity: A critical history of social media. Oxford University Press.
19. Woolard, K. A., & Schieffelin, B. B. (1994). Language ideology. Annual Review of Anthropology, 23, 55–82.
20. Zuboff, S. (2019). The age of surveillance capitalism. PublicAffairs.