Human-centered Artificial Intelligence for gender-inclusive language evolution*

Daniela Vellutino, Mafalda Ingenito and Giuliana Vitiello (Salerno)

Abstract

Achieving linguistic inclusivity in artificial intelligence systems cannot be accomplished by technological advancements alone. This paper underscores the necessity of a multidisciplinary and collaborative approach involving linguists, sociologists, software developers, and policy-makers to ensure that AI-driven language technologies are not only technically robust but also culturally sensitive and ethically sound. Equally crucial is the active engagement of end-users in the design and implementation of inclusive AI systems, fostering a sense of empowerment rather than imposition.

Promoting inclusive language through AI is therefore not merely a technical endeavor, but a shared social responsibility. Language that reflects and respects human diversity enhances communication and contributes to a more just and equitable society. The future of AI in the domain of language will depend on our collective capacity to design systems that are transparent, accountable, and responsive to the evolving needs of a pluralistic global community.

1 Introduction

Artificial Intelligence (AI) has transformed the way we interact with technology, including how we teach, learn, and process language, by enabling systems to learn, reason, and autonomously generate linguistic content (cf. Alaqlobi et al. 2024). However, these advancements also raise critical challenges related to fairness and inclusivity, as AI algorithms can inadvertently absorb and amplify gender biases embedded in training data, ultimately shaping communication in ways that mirror cultural and social stereotypes (cf. Güven et al. 2025).

From various research perspectives, studies unequivocally demonstrate that gender bias is present in AI datasets in general, and more specifically in training datasets used for Large Language Models (cf. Lu et al. 2020). These models are trained on vast amounts of data sourced from the Internet, which may contain discriminatory opinions (cf. UNESCO, IRCAI 2024).

The presence of gender-related stereotypes, often with a pejorative connotation towards women, is linguistically evident in grammatical asymmetries, which also manifest as semantic

* The authors share full scientific responsibility for the article as a whole. Daniela Vellutino is the author of Sections 1; 3, 3.1, 3.2, 3.3, and 3.4; Mafalda Ingenito is the author of Sections 2.1, 3.4.2, 3.4.3 and 3.4.4; Giuliana Vitiello is the author of Sections 2 and 4. Section 5 was co-authored.

disparities. As Cecilia Robustelli (2000) recalls, as early as the second half of the 19th century, Tennessee Claflin (1871) identified the phenomenon of semantic asymmetry through an analysis of the semantic field of honor: courtesan vs. courtier, free man vs. free woman. In the mid-1970s, Muriel Schulz (1975) defined this linguistic phenomenon as the "semantic derogation of women", referring to the shift in meaning that occurs when a masculine morphological form is transformed into its feminine counterpart. This linguistic phenomenon is widespread across multiple languages and was identified in Italian in the pioneering work of Sabatini/Mariani (1987).

Gender asymmetries in grammatical usage reinforce stereotypes by either attributing negative connotations to women and non-binary persons or by including them under the generic masculine, rendering them invisible (cf. Sczesny/Formanowicz/Moser, 2016). From a linguistic perspective, in languages such as Italian (cf. Vellutino 2018), where gender inflection is a defining typological feature based on the male/female binary opposition, it is crucial to promote language learning practices that counteract the dominance of masculine grammatical forms, which otherwise risk erasing women and non-binary individuals (cf. Thornton 2022).

Therefore, the use of linguistic datasets and metadata developed through approaches aimed at neutralizing gender bias and enhancing female and non-binary visibility is essential, particularly in relation to text type and communicative interaction in supervised learning. AI thus has the potential to be part of the solution in advancing gender equality within our societies.

To address this issue, the Human-Centered Artificial Intelligence (HCAI) paradigm advocates for an approach that places human well-being at the core of technological development. The goal is to design intelligent systems that are not only reliable and transparent but also promote ethical and inclusive language use (cf. Schmager/Pappas/Vassilakopoulou, 2025).

This work offers a critical synthesis of AI techniques for addressing gender bias in language, focusing on the intersection of linguistic theory and computational practice. Through an HCAI perspective, it reviews methods such as embedding analysis, co-occurrence metrics, and contextual models, highlighting debiasing strategies designed to foster fairness and inclusivity in Natural Language Processing (NLP). The contribution lies in proposing an interdisciplinary framework that connects technical solutions with ethical and cultural considerations.

2 Human-centered Artificial Intelligence

The concept of AI is intricate and multidimensional, encompassing a broad spectrum of forms and applications. Consequently, its definition remains intricate and subject to varying interpretations, to the extent that no universally accepted definition of the term exists (cf. Russell/ Norvig, 2016).

In general terms, AI can be described as the capability of an artificial system to perform tasks that traditionally require human intelligence. These tasks include learning, reasoning, problem-solving, perception, and, in some cases, creativity (cf. Shabbir/Anwer 2018).

A distinguishing feature of AI is its ability to autonomously learn from data: through training processes on large datasets, AI systems progressively enhance their performance, thereby reducing the need for direct human intervention. However, this very capacity for learning carries

with it a significant challenge: the algorithms, when trained on datasets that may contain implicit biases or distortions, risk the internalization and perpetuation of cultural, social, or gender-related prejudices (cf. Fournier-Tombs, Castets-Renard 2021). Consequently, the study of Artificial Intelligence cannot be disentangled from broader ethical considerations and the principles of transparency, to promote a responsible and inclusive use of these technologies.

The paradigm of Human-Centered Artificial Intelligence emerges within this framework. Shneiderman (2020) proposes that HCAI aims to augment and enhance human capabilities, ensuring that AI systems become more reliable, secure, and trustworthy. This approach entails a shift in perspective: the conception of AI as a substitute for human intelligence is replaced by the notion of it being a complement to, and a tool designed to support and amplify, human cognitive abilities. In this regard, it is imperative for AI to not only enhance its own performance but also to do so with greater consideration for the human and social context in which it operates. Adopting a human-centered approach has the potential to engender greater trust in AI technologies, rendering them more comprehensible and accessible to a broader range of users. This paradigm thus promotes continuous and transparent collaboration between humans and machines; wherein human intervention remains an essential element in the decision-making process.

Considering these considerations, Shneiderman (2020) introduces the concept of the "Second Copernican Revolution", a paradigm that redefines the relationship between humans and technology. To comprehend the magnitude of this transformation, it is instructive to recall the First Copernican Revolution, which occurred in the 16th century with the geocentric conception of the universe being overturned by Nicolaus Copernicus, thus demonstrating that the Earth did not occupy a central position but instead revolved around the Sun. This paradigm shift, far from being merely a scientific breakthrough, precipitated a profound cultural and philosophical transformation, redefining humanity's place in the cosmos.

In a similar vein, the Second Copernican Revolution proposed by Shneiderman (2020) posits a radical transformation in the interaction between humans and technology. While the primary objective of AI development has historically been the creation of autonomous systems capable of operating without human intervention, such as industrial robots or automated decision-making algorithms, this novel perspective restores human beings to the center of the technological system.

The transition from the "human in-the-loop" approach to the "AI in-the-loop" model marks a pivotal shift in the field. The former model relegated human involvement to a marginal role in autonomous systems' decision-making processes, whereas the latter integrates AI within human decision-making without, however, granting it ultimate authority over choices. In this new scenario, Artificial Intelligence serves as a tool to enhance human decision-making and analytical capabilities, rather than replace human critical judgment.

The shift from "human-in-the-loop" to "AI-in-the-loop" models redefines the balance between human and machine roles in collaborative systems. While the former centers human expertise in training and oversight, the latter embeds AI into decision-making workflows as a supportive tool, enhancing but not replacing human judgment (cf. Natarajan et al. 2025). This transition

reflects a move from automation to collaboration, with implications for bias, accountability, and ethical responsibility in sensitive domains.

Consequently, the concept of HCAI is not merely concerned with improving the efficiency of intelligent technologies but rather with promoting a more conscious and responsible use of AI, ensuring a synergistic interaction between humans and machines.

2.1 Principles of HCAI

As previously outlined, the design of HCAI extends beyond the development of advanced technologies to include the creation of systems that place human values, needs, and experiences at the core of development. This approach is grounded in a structured set of principles, ranging from ethical foundations to emotional intelligence, that ensure AI systems are inclusive, trustworthy, and aligned with societal expectations (cf. Pyae 2025).

Among these, the most relevant are:

- **Empathy**: An AI system must be capable of recognizing and responding to users' emotions, facilitating smoother and more natural interactions while improving usability, trust, and inclusivity. Although machines cannot experience emotions, they should be able to identify users' emotional states through the analysis of written and spoken language, as well as the recognition of facial expressions and micro expressions. Subsequently, the system must contextualize the detected emotion and generate an appropriate response that allows the user to feel understood and supported, for example, by modulating language tone or incorporating reassuring elements (cf. Zhu/Luo 2023; Cao et al. 2021).
- **Inclusivity**: AI technology design must account for human diversity in terms of gender, ethnicity, culture, age, and ability to ensure that AI serves as an accessible and beneficial tool for all. However, inclusivity is often undermined by multiple sources of bias, which go beyond training data alone. These include bias in data collection, annotation, model design, and even in evaluation practices (cf. Hovy/ Prabhumoye. 2021). To mitigate such risks and ensure a fair distribution of AI benefits, it is crucial to involve representatives of diverse social groups throughout the entire design and development process (cf. Anderson et al. 2024).
- **User-Centered Orientation***:* The development process of AI technologies must be guided by the needs, experiences, and values of end users. People should not be considered mere recipients of technology but rather active participants in its design. The objective is not only to create more powerful or faster AI but to develop systems that enhance quality of life, support decision-making processes, and are intuitive and accessible. An AI system that disregards user needs is bound to fail, regardless of its technical sophistication (cf. Usmani/ Happonen/Watada 2023).
- **Explainability**: The ability to make AI decision-making processes comprehensible is essential for fostering trust and ensuring informed use (cf. Ferrario/Loi 2022). Complex AI models, such as deep neural networks, are often perceived as "black boxes". Explainability is particularly crucial in fields where AI decisions have a direct impact on people's lives, such as healthcare or the judiciary. Key elements of this principle include the development of interpretable models, the provision of explanations tailored to users' expertise levels, and compliance with transparency regulations (cf. Lisboa et al. 2023).

- **Transparency and Trust***:* These two elements represent fundamental pillars for the safe adoption of AI technologies. Transparency entails understanding how AI systems function, what data they use, and which algorithms and logics they follow, while trust refers to users' confidence that AI systems are reliable, fair, secure, and under human control (cf. Schmidt/Biessmann/Teubner 2020).
- **Data Privacy***:* The protection and ethical management of personal data are central aspects of human-centered AI. As discussed by Martin/Zimmermann (2024), AI technologies challenge traditional privacy frameworks by automating data collection and processing, often making it harder for individuals to assess risks and retain control over their information. Ensuring informed consent and transparent data handling is thus essential to maintaining user trust and regulatory compliance.
- **Ethical Considerations***:* AI must be developed with ethical awareness, ensuring respect for human rights and dignity while actively countering biases in data and algorithms. Since AI systems reflect historical and social distortions, continuous oversight is essential to promote fairness and support inclusive, responsible technological interactions. As Safdar/Banja/Meltzer (2020) highlight, algorithmic decisions are shaped by the data and values embedded in their design, making neutrality a misconception. Ensuring fairness, transparency, and accountability is not optional but essential to prevent harm and build technologies that reflect social responsibility and moral integrity.

3 AI strategies for gender-fair tasks

This section aims to present various AI techniques that can be employed to detect gender bias in texts, as well as different approaches to making language more gender inclusive. NLP plays a crucial role in analyzing large volumes of text to identify linguistic patterns that contribute to gender bias. It is a subfield of artificial intelligence focused on enabling interaction between humans and computers through natural language, to develop algorithms that can understand, analyze, and generate human language (cf. Dande/Pund, 2023). Through this, it is possible to examine how language is used and to detect patterns that implicitly or explicitly discriminate based on gender.

But how does AI identify linguistic patterns that perpetuate gender bias? As Stanczak/Augenstein (2021) highlight, gender bias in language can emerge at both structural and contextual levels. Structural bias occurs when sentence constructions reflect stereotyped gender patterns, such as defaulting to masculine interpretations of neutral terms or explicitly marking gender. In contrast, contextual bias arises from lexical choices, tone, or framing, and requires background knowledge and interpretation to be identified. Detecting such bias is complex, as it involves both linguistic and extra-linguistic cues.

By analyzing the context in which a word is used, AI can detect implicit gender stereotypes encoded in language use, for example, it may find that words like *nurse* or *receptionist* are more closely aligned with female subjects, while words like *architect* or *warrior* align more with male subject, reflecting and reinforcing societal biases (cf. Bolukbasi et al. 2016). These patterns emerge from data because the language we use is deeply influenced by social norms and traditional gender roles.

Through the analysis of these patterns, AI can identify whether systematic trends reflecting gender bias exist, enabling informed actions to mitigate them.

The following subsection explores the main techniques utilized, including:

- Word Embedding Association Test;
- Co-occurrence Analysis;
- Contextual Models.

3.1 Word Embedding Association Test (WEAT)

The Word Embedding Association Test (WEAT) is a powerful method, designed to quantify and measure implicit biases present in word embeddings, numerical vectors that represent the semantic meanings of words in AI models (cf. Caliskan/Bryson/Narayanan 2017).

To fully understand how WEAT functions, it is essential to grasp the concept of word embeddings. When training AI models for NLP, words are transformed into numerical vectors within a multidimensional space, where words with similar meanings are positioned closer together (cf. Almeida/Xexéo 2019). WEAT helps measure the strength with which certain words, for example those related to gender or ethnicity, are associated with stereotypical concepts (cf. Du/Wu/Lan 2019; Van Loon et al. 2022).

Consider two sets of words representing gender, such as *man*, *father*, and *he* for the male category, and *woman*, *mother*, and *she* for the female category. WEAT examines whether implicit associations exist between these groups. These gendered words are then compared with two sets of neutral words representing concepts or professions, such as *scientist*, *engineer*, *nurse*, and *assistant*. Subsequently, the semantic distance between the words is measured, indicating how close or far the word vectors of the reference groups are from the attribute groups. This comparison reveals whether implicit associations exist between specific concepts and a particular gender. Finally, WEAT generates an association score, which indicates the strength of the relationship between the target words (e. g., gender-related terms) and attribute words (e. g., personal traits, professions, etc.). This score provides a quantitative measure of bias, enabling researchers to identify and understand underlying stereotypes embedded within language models.

3.2 Co-occurrence analysis

The co-occurrence analysis technique examines how words appear together within a given text or dataset. By analyzing these associations, it is possible to uncover hidden patterns that reflect biases embedded in language. In the context of AI, this technique is particularly useful for exploring how words related to certain concepts tend to co-occur, revealing implicit stereotypes and cultural norms (cf. Sedighi 2016).

But how does this approach work? First, a large corpus of texts is required. This dataset can consist of newspaper articles, social media posts, academic papers, or other written sources that reflect societal language use. Once the corpus is compiled, the next step is to calculate how frequently certain words appear together. This analysis can be conducted at different levels, such as within a sentence, a paragraph, or an entire document, depending on the research objective.

After collecting co-occurrence data, the strength of the association between words can be measured using statistical metrics such as association scores or the chi-square test. These metrics help determine how strong and significant the connections between words are. For example, co-occurrence analysis might reveal that words like *leader* or *executive* frequently appear with male pronouns, while words like *assistant* or *caregiver* are more commonly associated with female pronouns. Such patterns reflect societal stereotypes and can contribute to reinforcing gender biases in language.

One of the primary advantages of co-occurrence analysis is its simplicity. It effectively identifies repetitive word associations without requiring complex linguistic models, making it a quick and efficient method for revealing hidden linguistic patterns. However, it has an inability to account for the context in which words appear. Two words may co-occur in the same text, but their meanings can vary significantly depending on the context. For example, the word pair *doctor* and *she* might appear in a sentence discussing gender stereotypes rather than reflecting an actual societal association.

3.3 Contextual models

In language, the meaning of a word is heavily dependent on its context, the other words that accompany it. For example, the word "bank" can mean a financial institution or the side of a river. The correct interpretation can only be determined by examining the surrounding words. Contextual models help AI understand these nuances, significantly enhancing its ability to interpret human language (cf. Naseem et al. 2021).

To understand how contextual models function, it is essential to examine their architecture and information-processing methods. One of the most influential models in this field is BERT (Bidirectional Encoder Representations from Transformers, cf. Gardazi et al. 2025). Unlike earlier models that processed words sequentially, BERT and similar models, such as GPT (Generative Pre-trained Transformer) (cf. Topal/Bas/van Heerden 2021), analyze an entire sentence or paragraph simultaneously, considering each word about all the others.

For example, if BERT encounters the sentence: *The bank is located along the river*, it uses the context of *river* to understand that *bank* refers to the riverbank rather than a financial institution. This bidirectional approach enables a much richer and more accurate understanding of language compared to traditional models, which would have interpreted each word separately without considering its broader meaning.

A key feature of contextual models is their ability to learn from both directions of a sentence, capturing long-term dependencies between words. This means that BERT can interpret both the preceding and following context of a word, leading to a more comprehensive understanding of the language. For instance, in the sentence *She went to the bank to watch the boats sail by*, the model can use the phrase *boats sail* to accurately infer that *bank* refers to a riverside location, not a financial institution.

This bidirectional processing is powered by Transformers (cf. Lin et al. 2022), a neural network architecture that allows the model to focus on different parts of the input simultaneously. Transformers use a mechanism called self-attention, which enables the model to weigh the

importance of each word relative to others in the sentence. This results in a dynamic representation of words, influenced by the complete context in which they appear.

3.4 Debiasing

At this point, let us introduce a set of techniques known as Debiasing (cf. Meade/Poole-Dayan/Reddy 2021). Debiasing refers to the process of removing biases from artificial intelligence models, particularly linguistic ones. It aims to correct models that reflect gender stereotypes or other forms of bias present in training data (cf. Sokolová et al. 2024).

Why is debiasing so crucial in our case? The answer is simple: without debiasing, AI models risk perpetuating and amplifying existing inequalities.

Several debiasing techniques have been developed to render texts fair and gender-inclusive, including:

- Neutralization and Equalization
- Fine-tuning on balanced data
- Sequence-to-sequence debiasing
- Data augmentation for debiasing
- Bias attenuation through attention mechanisms

3.4.1 Neutralization and equalization

To address gender bias in language models, two complementary debiasing techniques are commonly employed: neutralization and equalization. Both approaches operate on word embeddings, but they target different aspects of biased associations.

Neutralization focuses on removing unintended gender associations from words that should be inherently gender neutral. For instance, terms such as *doctor*, *scientist*, or *leader* should not carry implicit gendered meanings. However, due to historical and societal biases embedded in training corpora, such terms are often disproportionately associated with male pronouns or characteristics. The neutralization process aims to project these words away from the "gender direction" in the embedding space, thereby eliminating or minimizing their alignment with male or female features. The result is a set of word vectors that are more gender-neutral, helping to reduce the reinforcement of stereotypes in downstream tasks.

Equalization, on the other hand, seeks to ensure that word pairs representing opposing genders, such as *man* and *woman* or *father* and *mother*, are treated symmetrically. This technique is applied when such pairs should be semantically equivalent but are unequally represented in the training data. Equalization adjusts their embeddings to reflect an equal relationship, aligning their associations with traits and concepts in a balanced manner. For example, if the word *man* is more closely associated with *strong*, *intelligent*, or *rational* while *woman* is linked to *emotional* or *weak,* the model risks perpetuating harmful stereotypes.

Equalization intervenes by modifying the vectors to balance these associations, ensuring that both terms are represented equitably in the model's output.

3.4.2 Fine-tuning

Fine-tuning refers to the process of further optimizing a pre-trained model, often to improve its performance on specific tasks or to adapt it to a new domain (cf. Zhang/Li/Liu 2024). An essential consideration when discussing fine-tuning is the quality and balance of the data used in this phase. Why is it so important to work with balanced data during fine-tuning? AI models learn to make decisions and predictions based on patterns found in the training data. If the data is unbalanced, the model tends to favor the more represented classes or categories, neglecting those that are less present.

Fine-tuning on balanced data begins with identifying and gathering a more representative dataset. This means ensuring that all relevant classes, categories, or groups are equally represented in the data used for training and fine-tuning.

Once the balanced dataset is created, fine-tuning can proceed. The model, which was previously trained on a large corpus of data, is now fine-tuned using this new, balanced dataset. This phase is crucial to improving the model's accuracy in specific contexts and reducing any biases that may have emerged during the initial training phase.

Models like BERT are trained on vast amounts of textual data from sources such as articles, books, or social media posts. However, even these datasets can be biased, reflecting social, cultural, or gender biases. For example, if a linguistic model is primarily trained on texts written by male authors, it may learn to favor certain expressions or linguistic styles that reflect a male perspective, ignoring or undervaluing other viewpoints. Fine-tuning on a more balanced dataset, which includes texts written by a broader range of authors, ensures that the model can respond more impartially and inclusively, treating different perspectives with equal importance.

Fine-tuning on balanced datasets offers several crucial advantages for enhancing the quality and fairness of AI models. First, it helps mitigate implicit bias by ensuring equal representation across all classes or categories, an essential requirement for applications in sensitive domains like healthcare, education, and the justice system.

Additionally, balanced fine-tuning improves the model's ability to generalize to unseen data. When no class is overrepresented, the model is more likely to perform reliably across diverse scenarios and user groups, thereby reducing the risk of biased or inaccurate predictions.

3.4.3 Data Augmentation

"Data Augmentation" refers to the technique of expanding the training dataset by generating new examples from existing data (cf. Maharana/Mondal/Nemade 2022). This approach is particularly valuable in debiasing, as it allows for balancing the classes and categories within the data, correcting implicit biases. Data augmentation is especially useful in domains where AI models need to make accurate predictions based on patterns identified in data, as it mitigates biases or imbalances by increasing the diversity and representativeness of the data, thereby improving the model's performance and impartiality.

Consider an AI model used for text classification or generation. If the model were trained on data where technical occupations are predominantly associated with men, we could generate new examples where the same occupations are associated with women. For example,

transforming a sentence like **He is an engineer** into **She is an engineer** would balance the gender representation within the dataset. This process helps the model learn to treat men and women equally in its outputs. The benefits of this technique include balancing classes, increasing data diversity, and improving model accuracy.

3.4.4 Sequence-to-Sequence

Sequence-to-Sequence (Seq2Seq) Debiasing is a specialized technique applied to sequence-to-sequence models, which are commonly used in natural language processing tasks such as machine translation, chatbots, and text generation (cf. Neubig 2017). These models function by taking a sequence of inputs and generating a corresponding sequence of outputs. They typically employ an encoder-decoder architecture, where the encoder processes the input (which can be a sentence or a sequence of words) and converts it into an internal, vector-based representation. This representation is then passed to the decoder, which generates the desired output sequence.

For instance, in machine translation, a sequence of words in one language is converted into a cor-responding sequence in another language (cf. Wang et al. 2022). However, even in these models, implicit biases in the training data may reflect social biases, such as gender stereotypes. Bias in these models poses a significant issue. In machine translation, gender-neutral pronouns in source languages, like English, may be translated into gendered pronouns in languages such as Spanish or French, depending on the associations the model has learned. If the training data contains more references to men in professional roles than to women, the model is likely to generate male pronouns, even when the context does not specify gender.

Seq2Seq Debiasing aims to address these biases by implementing a set of strategies to ensure that the model produces more equitable and inclusive results.

These strategies include:

- Correcting the training data: Balancing the training data so that job roles, character traits, and pronouns are more equally represented across genders. This ensures that the model learns from data that reflects greater gender equality.
- Modifying the model structure: In some cases, the internal representations of the Seq2Seq algorithm can be adjusted to neutralize tendencies, such as the propensity to generate masculine pronouns when gender is not explicitly indicated in the input.
- Re-evaluating output sequences: After generating an output sequence, filters or control algorithms can be applied to ensure that the results are free from gender bias. This final step allows real-time monitoring of the generated text to correct any biases produced by the model.

4 Bias Mitigation through Attention Mechanisms

The final technique we will discuss is **bias mitigation through attention mechanisms**. In recent years, attention mechanisms have revolutionized the field of NLP and neural networks by enhancing the ability of models to focus on the relevant parts of an input sequence. These mechanisms have not only improved model performance but have also emerged as powerful tools for mitigating bias.

Before delving into bias mitigation, it's important to understand how attention mechanisms work. Attention mechanisms are key components in deep learning models such as the Transformer architecture, which underpins advanced models like BERT. These mechanisms allow the model to "focus" on specific parts of the input when generating an output sequence. Rather than treating each word or element of the input equally, the model can prioritize certain words or phrases that are most relevant to the task (cf. Vaswani et al. 2017). For example, in machine translation, attention mechanisms help the model identify which words in the source sentence are most important for accurately translating each word into the target language. This technique has been shown to be highly effective in improving the quality of translations and the model's understanding of linguistic context. However, the true potential of attention mechanisms lies in their ability to mitigate bias. Due to their adaptive nature, these mechanisms can be leveraged to identify and correct biased patterns in the training data, allowing the model to focus on more neutral or equitable information. Bias mitigation through attention mechanisms occurs by altering how the model assigns attention to different parts of the input sequence. Since attention mechanisms enable the model to focus on specific elements, we can control which parts of the text receive attention, thus diminishing the weight of information that reinforces biases and stereotypes. One way to implement bias mitigation is by adjusting attention weights. If the model tends to focus more on terms that perpetuate gender stereotypes, we can reduce the attention given to these terms and increase the focus on more neutral elements. This approach allows the model to make predictions or generate responses that are less influenced by implicit biases in the data.

5 Conclusion

Artificial Intelligence is rapidly transforming the way we interact with language, providing increasingly sophisticated tools for text processing and generation. However, this progress comes with challenges, particularly in preventing the reinforcement and amplification of gender biases embedded in training data. The primary goal is not just to make AI more efficient but to ensure that it operates ethically, fostering fairness and inclusivity in language.

Through the analysis of various methodologies, such as the Word Embedding Association Test, co-occurrence analysis, and contextual models, it has become evident that AI can serve as a powerful tool for detecting and quantifying biases in text. At the same time, debiasing techniques, including neutralization and equalization of semantic vectors, fine-tuning on balanced data, and the use of attention mechanisms, show that it is possible to mitigate these distortions and promote a more equitable linguistic landscape.

Yet, it is essential to move beyond optimistic narratives. Saying that eliminating bias is complex is not enough: research has shown that many debiasing methods offer only partial mitigation and may obscure rather than resolve underlying structural issues.

Crucially, AI alone cannot ensure linguistic inclusivity. Technological solutions must be accompanied by the active engagement of linguists, sociologists, developers, and policymakers, to ensure that models are not only technically sound but also culturally and ethically robust. Moreover, user involvement in the design and deployment of inclusive AI systems is essential to foster trust and to align these tools with real-world expectations and values.

Ultimately, building a more inclusive language through AI is not just a technical undertaking – it is a social and ethical responsibility. A linguistic system that reflects and values diversity enhances communication and actively contributes to a more just society. The future of AI in language depends on our collective ability to design systems that are transparent, accountable, and responsive to the evolving needs of a pluralistic world (cf. Gonen/Goldberg 2019).

It is crucial to recognize that AI alone cannot resolve the challenge of linguistic inclusivity. Techno-logical solutions must be supported by the intentional and collaborative efforts of linguists, sociologists, developers, and policymakers to ensure that innovations are not only technically sophisticated but also culturally and ethically sound. Equally vital is the active participation of users in the design and adoption of inclusive AI systems, so that these technologies are perceived as empowering tools rather than externally imposed constraints.

Ultimately, promoting inclusive language through AI is not solely a technical task – it is a shared social responsibility. Language that embraces and represents diversity not only enhances communication but also fosters a more just and equitable society. The future of AI in the realm of language will depend on our collective capacity to build systems that are transparent, accountable, and responsive to the needs of a continually evolving world.

References

Alaqlobi, Obied et al. (2024): "Artificial intelligence in applied linguistics: a content analysis and future prospects". *Cogent Arts & Humanities* 11/1: 2382422.

Almeida, Felipe/Xexéo, Geraldo (2019): "Word embeddings: A survey". *arXiv preprint* arXiv:1901.09069 [05.01.2026].

Anderson, Andrew et al. (2024): "Measuring User Experience Inclusivity in Human-AI Interaction via Five User Problem-Solving Styles". *ACM Transactions on Interactive Intelligent Systems* 14/3: 1–90.

Bolukbasi, Tolga et al. (2016): "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". doi.org/10.48550/arXiv.1607.06520

Caliskan, Aylin/Bryson, Joanna J./Narayanan, Arvind (2017): "Semantics derived automatically from language corpora contain human-like biases". *Science* 356/6334: 183–186.

Claflin, Tennessee Celeste (1871): *Constitutional Equality a Right of Woman; a consideration of the various relations which she sustains as a necessary part of the body of society, etc. [With a portrait.]*. New York: Woodhull, Claflin & Company.

Cao, Siqi et al. (2021): "Can AI detect pain and express pain empathy? A review from emotion recognition and a human-centered AI perspective". doi.org/10.48550/arXiv.2110.04249.

Dande, Abhay A./Pund, D. (2023): "A review study on applications of natural language processing". *International Journal of Scientific Research in Science, Engineering and Technology* 10/2: 122–126.

Du, Yupei/Wu, Yuanbin/Lan, Man (2019): "Exploring human gender stereotypes with word association test". *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP): 6133–6143.

Ferrario, Andrea/Loi, Michele (2022): "How explainability contributes to trust in AI". *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. New York, ACM: 1457–1466. doi.org/10.1145/3531146.3533202

Fournier-Tombs, Eleonore/Castets-Renard, Céline (2021): "Algorithms and the Propagation of Gendered Cultural Norms". In: Guèvremont, Véronique/ Brin, Colette (eds.): *IA, Culture et Médias.* Presses de l'Université de Laval. dx.doi.org/10.2139/ssrn.3980113.

Gardazi, Nadia Mushtaq et al. (2025): "BERT applications in natural language processing: a review". *Artificial Intelligence Review* 58/6: 1–49.

Gonen, Hila/Goldberg, Yoav (2019): "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them". *arXiv preprint*. arXiv:1903.03862 [05.01.2026].

Güven, Çiçek et al. (2025): "AI in Support of Diversity and Inclusion". arXiv:2501.09534 [05.01.2026].

Hovy, Dirk/Prabhumoye, Shrimai (2021): "Five sources of bias in natural language processing". *Language and Linguistics Compass* 15/8: e12432.

Lin, Tianyang et al. (2022): "A survey of transformers". *AI Open* 3: 111–132.

Lisboa, Paulo J. G. et al. (2023): "The coming of age of interpretable and explainable machine learning models". *Neurocomputing* 535: 25–39.

Lu, Kaiji et al. (2020): "Gender bias in neural natural language processing". In: Nigam, Vivek et al. (eds.): *Logic, Language, and Security. Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*. Berlin, Springer: 189–202.

Maharana, Kiran/Mondal, Surajit/Nemade, Bhushankumar (2022): "A review: Data pre-processing and data augmentation techniques". *Global Transitions Proceedings* 3/1: 91–99.

Martin, Kelly D./Zimmermann, Johanna (2024): "Artificial intelligence and its implications for data privacy". *Current Opinion in Psychology*: 101829.

Meade, Nicholas/Poole-Dayan, Elinor/Reddy, Siva (2021): "An empirical survey of the effectiveness of debiasing techniques for pre-trained language models". *arXiv preprint*. arXiv:2110.08527[05.01.2026].

Natarajan, Sriraam et al. (2025): "Human-in-the-loop or AI-in-the-loop? Automate or Collaborate?". *Proceedings of the AAAI Conference on Artificial Intelligence* 39/27: 28594–28600.

Naseem, Usman et al. (2021): "A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models". *Transactions on Asian and Low-Resource Language Information Processing* 20/5: 1–35.

Neubig, Graham (2017): "Neural machine translation and sequence-to-sequence models: A tutorial". *arXiv preprint*. arXiv:1703.01619 [05.01.2026].

Pyae, Aung (2025): "What is Human-Centeredness in Human-Centered AI? Development of Human-Centeredness Framework and AI Practitioners' Perspectives". arXiv:2502.03293 [05.01.2026].

Robustelli, Cecilia et al. (2000): "Lingua e identità di genere. Problemi attuali nell'italiano". *Studi italiani di linguistica teorica e applicata* 29: 507–527.

Russell, Stuart J./Norvig, Peter (2016): *Artificial Intelligence: A Moßdern Approach*. Harlow: Pearson.

Sabatini, Alma/Mariani, Marcella (1987): *Il sessismo nella lingua italiana*. Roma: Presidenza del Consiglio dei ministri, Direzione Generale delle Informazioni.

Safdar, Nabile M./Banja, John D./Meltzer, Carolyn C. (2020): "Ethical considerations in artificial intelligence". *European Journal of Radiology* 122: 108768.

Schulz, Muriel (1975): "The semantic derogation of women". In: Thorne, Barrie/Henley, Nancy (eds.): *Language and Sex: Difference and Dominance*. Rowley, MA, Newbury House: 134–147.

Schmager, Stefan/Pappas, Ilias/Vassilakopoulou, Polyxeni (2025): "Understanding Human-Centred AI: a review of its defining elements and a research agenda". *Behaviour & Information Technology* 44/15: 3771–3810. doi.org/10.1080/0144929X.2024.2448719.

Schmidt, Philipp/Biessmann, Felix/Teubner, Timm (2020): "Transparency and trust in artificial intelligence systems". *Journal of Decision Systems* 29/4: 260–278.

Sczesny, Sabine/Formanowicz, Magda/Moser, Franziska (2016): "Can gender-fair language reduce gender stereotyping and discrimination?". *Frontiers in Psychology* 7: 154379.

Shabbir, Jahanzaib/Anwer, Tarique (2018): "Artificial intelligence and its role in near future". arXiv:1804.01396 [05.01.2026].

Sedighi, Mehri (2016): "Application of word co-occurrence analysis method in mapping of the scientific fields (case study: the field of Informetrics)". *Library Review* 65/1–2: 52–64.

Shneiderman, Ben (2020): "Human-centered artificial intelligence: Three fresh ideas". *AIS Transactions on Human-Computer Interaction* 12/3: 109–124.

Sokolová, Zuzana et al. (2024): "Measuring and mitigating stereotype bias in language models: An overview of debiasing techniques". Muštra, Mario/Vuković, Josip/ Bozek, Jelena (eds.): *Proceedings of the 2024 International Symposium* ELMAR. Zadar, Elmar: 241–246. researchgate.net/publication/384650376_Measuring_and_Mitigating_Stereotype_Bias_in_Language_Models_An_Overview_of_Debiasing_Techniques [20.02.2026].

Stanczak, Karolina/Augenstein, Isabelle (2021): "A survey on gender bias in natural language processing". *arXiv preprint*. arXiv:2112.14168 [05.01.2026].

Thornton, Anna M. (2022): "Genere e igiene verbale: l'uso di forme con in italiano". *Annali del Dipartimento di Studi Letterari, Linguistici e Comparati. Sezione linguistica* 11: 11–54.

Topal, M. Onat/Bas, Anil/van Heerden, Imke (2021): "Exploring transformers in natural language generation: GPT, BERT, and XLNet". *arXiv preprint*. arXiv:2102.08036 [05.01.2026].

Usmani, Usman Ahmad/Happonen, Ari/Watada, Junzo (2023): "Human-centered artificial intelligence: Designing for user empowerment and ethical considerations". *Proceedings of the 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA 2023)*. Piscataway/NJ, IEEE: 1–7. doi.org/10.1109/HORA58378.2023.10156761

UNESCO, IRCAI (2024): "Challenging systematic prejudices: an Investigation into Gender Bias in Large Language Models". *International Research Centre on Artificial Intelligence* 5. unesdoc.unesco.org/ark:/48223/pf0000388971 [09.06.2025].

Van Loon, Austin et al. (2022): "Negative associations in word embeddings predict anti-Black bias across regions – but only via name frequency". *Proceedings of the International AAAI Conference on Web and Social Media* 16: 1419–1424.

Vaswani, Ashish et al. (2017): "Attention is all you need". In: Guyon, Isabelle et al. (eds.): *Advances in Neural Information Processing Systems 30. NIPS 2017*. proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf [25.03.2026].

Vellutino, Daniela (2018): *L'italiano istituzionale per la comunicazione pubblica*. Bologna: Il Mulino.

Wang, Haifeng et al. (2022): "Progress in machine translation". *Engineering* 18: 143–153.

Zhu, Qihao/Luo, Jianxi (2023): "Toward artificial empathy for human-centered design: A framework". doi.org/10.48550/arXiv.2303.10583

Zhang, Yulin/Li, Yanhua/Liu, Junhan (2024): "Unified efficient fine-tuning techniques for open source large language models". Preprint (Version 1). doi.org/10.21203/rs.3.rs-4660140/v1.