Preprints.org

Review

Not peer-reviewed version

Rethinking Research on Stereotypes: An Analysis through Social Psychological and Computational Perspectives

Kaustubh Shivshankar Shejole * and Pushpak Bhattacharyya

Posted Date: 12 January 2026

doi: 10.20944/preprints202601.0815.v1

Keywords: stereotypes; social psychology; computational approaches; bias; responsible AI; computational social science; language/cultural bias analysis; sociolinguistics; NLP tools for social analysis; hate-speech detection; bias/toxicity; misinformation detection and analysis

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Review

Rethinking Research on Stereotypes: An Analysis Through Social Psychological and Computational Perspectives

Kaustubh Shivshankar Shejole * and Pushpak Bhattacharyya

Computation for Indian Language Technology (CFILT), Department of Computer Science and Engineering, Indian Institute of Technology Bombay, Mumbai, India

* Correspondence: kaustubhshejole@cse.iitb.ac.in

Abstract

Stereotypes are very harmful social constructs shaping human perception and behavior. Recent work shows that large language models (LLMs) may inherit and amplify such social harms. However, most existing research often focuses on stereotypical biases and overlooks stereotypes and the rich social-psychological literature on them, resulting in resource wastage and slowed progress in stereotype research. We argue that meaningful progress in mitigating stereotypes in LLMs requires tighter integration between social psychology and computational research. To address this gap, we review core social-psychological theories and frameworks and analyze their computational operationalization, highlighting substantial open opportunities. We also analyze computational progress across media narratives, body imaging, and multilingual, multicultural, and multimodal contexts, identifying key gaps and limitations in each domain. We also present a unified analysis of challenges in stereotype research. We further discuss implications for responsible AI, highlighting stereotypes as a root source of downstream harms, and briefly examine the limitations of current mitigation approaches along with potential improvements via explainability and interpretability. We frame stereotypes in AI as socio-technical phenomena and urge further research in responsible AI, informed by the perspectives and future directions presented in this paper.

Keywords: stereotypes; social psychology; computational approaches; bias; responsible AI; computational social science; language/cultural bias analysis; sociolinguistics; NLP tools for social analysis; hate-speech detection; bias/toxicity; misinformation detection and analysis

1. Introduction

Humans are quite good at identifying patterns and forming clusters. They naturally construct conceptual groupings based on shared features, even under uncertainty [1,2], though this process does not always follow strictly logical rules [3]. Humans tend to classify those similar to themselves as the "in-group" and those perceived as different as the "out-group" [4–7]. Theories such as the similarity–attraction hypothesis [8] and social identity theory [9] suggest that people are more attracted to others who share similar attitudes, values, and traits (i.e., in-groups). This gives rise to in-group favoritism and can also produce varied emotional responses toward out-groups, such as hate, pity, or respect, as studied by [4,7,10–12]. These feelings toward out-groups are gradually translated into thoughts, which then solidify into beliefs: what we call ***"stereotypes"***.

The human brain is evolutionarily tuned to respond rapidly to stimuli perceived as critical for survival, such as predators or fire, thereby prioritizing System 1 processing[1] [14–17]. Social competition and interpersonal relationships likewise constitute fundamental survival-relevant contexts [18–20],

[1] System 1 refers to fast, automatic, intuitive, and emotion-driven cognition, in contrast to System 2, which is slower, deliberate, and analytical [13].

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

and therefore tend to elicit fast responses governed by System 1. The origin of stereotypes can thus be attributed to System 1 thinking and the cognitive distinctions between "in-groups" and "out-groups".

LLMs are increasingly adopted across a wide range of domains, ranging from educational applications such as teaching assistants [21] to medical settings, including clinical report generation [22], to mention a few; their societal impact continues to expand. Recent work shows that LLMs inherit and sometimes amplify these stereotypes as they learn them from their large-scale pre-training corpora [23–25]. To mitigate these challenges, research has initially focused on assessing bias in LLMs by exploiting various tasks, including Natural Language Generation (NLG) [26,27], counterfactual reasoning [28,29], Natural Language Inference (NLI) [30], Question Answering [31,32], and prompt completion [33,34]. Gallegos et al. [35] provides a detailed analysis of dataset and metrics designed for assessing bias in LLMs. But fundamentally, bias is a different concept from stereotypes, and confusing biases with stereotypes can give rise to inefficient benchmarks, resulting in substantial resource waste [36]. These concepts are well-studied in social psychology; however, only a few papers draw on social-psychological insights, limiting progress in this domain. Stereotypes are the primary origin of inter-group relations and should therefore be studied separately to understand their effect in Responsible AI. For example, Tomar et al. [37] found that incorporating stereotype detection can improve bias detection accuracy. Thus, stereotypes hold considerable potential, which, if systematically explored, can significantly advance Responsible AI research.

To address this gap, this paper focuses primarily on stereotype research and analyzes it from social-psychological and computational perspectives. Our contributions are:

1. A systematic review of social-psychological theories and frameworks on stereotypes that will guide future computational research (Section 2). We also review the computational operationalization of these frameworks and theories, highlighting open opportunities. We analyze computational progress and gaps across domains such as narrative, media, and body imaging, and provide future directions (Section 3).
2. A multimodal, linguistic, and geographic analysis of stereotype research, identifying key gaps and underexplored requirements (Section 4).
3. A unified analysis of challenges in stereotype research by integrating social-psychological and computational perspectives (Section 5).
4. An analysis of implications for Responsible AI, framing stereotypes as foundational to downstream harms, and briefly examining existing mitigation approaches' failures, while suggesting potential improvements through explainability and interpretability (Section 6).

2. Social Psychological Perspectives on Stereotypes

In this section, we review key social psychological theories (Section 2.1) and frameworks (Section 2.2) on the formation, structure, and function of stereotypes.

2.1. Foundational Theories

1. *Similarity–Attraction and Social Identity Theory:* As discussed in the introduction, similarity-attraction theory [8] and Social Identity Theory [9] posit systematic *in-group* favoritism, whereby individuals favor in-groups over out-groups to enhance self-esteem [10]. Self-esteem comprises personal and social identity, the latter derived from group memberships based on attributes such as nationality or age. According to Social Identity Theory, threats to self-esteem intensify in-group favoritism, which in turn restores self-worth, a prediction supported empirically [38,39]. From this perspective, stereotypes function as mechanisms for self-esteem maintenance, emerging through in-group favoritism and out-group derogation when out-groups are perceived as threatening, thereby conceptualizing stereotypes as ***self-esteem protectors***.
2. *Social Role Theory:* This theory [40] focuses on socialization processes and posits that stereotypes are shaped by the social roles people occupy, such as lower-status versus higher-status jobs. Media plays a direct role in shaping stereotypes, often without individuals being consciously

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

aware of its influence [41]. In particular, media representations strongly affect body image by promoting stereotypical ideals, such as muscular and lean bodies for males, and fashionable, thin bodies for females [42,43]. Social Role Theory is closely related to Social Learning Theory [44], as both emphasize learning through observation and social reinforcement. These theories conceptualize stereotypes as ***social representations*** representing existing social roles.

3. *Social Categorization Theory:* This theory states that group-based perception is as fundamental as individual-based perception [45]. It argues that stereotyping and categorization are the two central components of perception. It states that both the process of stereotyping and the content of stereotypes are fluid and dynamic, varying across social contexts. Social context determines the nature of *self–other* comparisons and shapes how group boundaries are constructed. It considers that stereotypes reflect the emergent properties of social groups. It conceptualizes stereotypes as ***psychologically valid representations*** [46], grounded in group-based cognition.
4. *Theories Discussing Social Cognition:* Social cognition–based theories [47–50] conceptualize stereotyping as a "necessary evil," arising from the human cognitive need for simplicity and order. These theories view stereotypes as cognitive functions that simplify the complexity of the social world through implicit and often automatic processes. These theories conceptualize stereotypes as ***cognitive schemas*** structuring perception.
5. *Social Justification Theory:* This theory [51–53] states that holding negative stereotypes of another group may serve not only an ego-protective and group-protective function, but also a system-justifying function. It argues that when status hierarchies relegate groups to relative positions of inferiority and superiority, members of disadvantaged groups may themselves come to hold negative beliefs about their own groups in the service of a larger system in which social groups are hierarchically arranged [54]. This theory states that stereotypes can be considered as reinforcing the ideology of dominant groups, which may even be endorsed by disadvantaged groups themselves. It considers stereotypes as ***ideological representations***.
6. *Discursive Philosophy of Categorization:* The previous approaches consider categorization as highly functional and adaptive, and are largely grounded in a realist epistemology (i.e., the assumption that reality can be understood through facts or reason). Discursive philosophy challenges this realist epistemology. It does not treat social categories as rigid internal entities used inflexibly; instead, it is concerned with how people discursively construct social categories. It examines how these constructions produce subjectivities for both the self and those defined as the "Other." Wetherell and Potter [55] states that people are often inconsistent and highly context-dependent in articulating their beliefs. According to this perspective, stereotypes are relatively stable, shared, and identifiable, yet emerge through discourse rather than internal cognition. Similarly, Edwards [56] conceptualize stereotypes and categorization as ***discursive constructions*** rather than cognitive processes [46].
7. *Intersectionality Theory:* Recent work [57–59] emphasizes that social identities such as race, gender, and ethnicity interact rather than operate independently. From this perspective, stereotypes are not isolated constructs but emerge through the intersection of multiple identity dimensions, producing distinct and context-dependent forms of discrimination (e.g., experiences specific to Asian American women). Intersectionality thus frames stereotypes as ***relational and co-constructed structures*** across social categories.

Though various theories conceptualize stereotypes in different ways, all of them acknowledge the harms associated with them.

2.2. Major Frameworks

1. *Stereotype Content Model (SCM):* The SCM proposes that group stereotypes are structured along two fundamental dimensions: warmth (perceived intent) and competence (perceived ability) [7]. Warmth judgments are shaped primarily by perceived competition, while competence judgments reflect perceived status. These dimensions yield four canonical stereotype profiles:

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

admiration (high warmth, high competence; e.g., ingroups), pity (high warmth, low competence; e.g., the elderly or people with disabilities), envy (low warmth, high competence; e.g., high-status outgroups), and contempt (low warmth, low competence; e.g., stigmatized groups). Each quadrant is associated with distinct emotional and behavioral tendencies, ranging from active facilitation to active harm, enabling the SCM to predict real-world social behaviors such as inclusion, neglect, or discrimination [7,60].

2. *Agency–Beliefs–Communion (ABC) Model:* The ABC model[2] [62] reframes stereotype content by positing that social perception is fundamentally organized around *Agency* (socioeconomic power) and *Beliefs* (ideological orientation), rather than the warmth-competence dimensions central to the SCM. Developed as a critique of SCM, it challenges its theory-driven structure and reliance on predefined social groups, which may limit the discovery of naturally salient dimensions. Adopting a bottom-up approach, the ABC model shows that *Communion* (including warmth and morality) is not a primary dimension but an emergent construct arising from combinations of Agency and Beliefs. Empirical evidence across multiple studies indicates that spontaneous group categorization aligns most strongly with these two dimensions: Agency shapes power-related judgments, while Beliefs capture ideological alignment. Notably, groups at extreme levels of Agency are perceived as low in communion, whereas moderate Agency is associated with higher communal attributions, suggesting that warmth-based judgments are secondary rather than foundational.
3. *Dual-Perspective Model:* The SCM proposed by Fiske et al. [7] considers competence as Agency (A) and warmth as Communion (C). Abele et al. [63] observed that A and C contain multiple components; for example, masculinity (e.g., "assertive" or "decisive") is also part of Agency, while morality (e.g., "fair," "honest") is part of Communion. They proposed a facet model that differentiates A into assertiveness (AA) and competence (AC), and C into warmth (CW) and morality (CM), and reported a good model fit.
4. *Five-Tuple Framework:* Both Davani et al. [64] and Shejole and Bhattacharyya [36] converge on a five-tuple framework for characterizing stereotypes, consisting of the *target group* (T), *relationship characteristics* (R), *associated attributes* (A), the *perceiving group or community* in which the stereotype is held (C), and the *context or time interval* (I) in which it emerges. Both works emphasize that stereotypes are inherently dynamic, varying across social groups and evolving over time, rather than being static representations. This perspective aligns with earlier social psychological theories highlighting the context-dependent and socially constructed nature of stereotyping [45]. This framework is particularly valuable for computational modeling of stereotypes, as it enables the integration of diverse methodological approaches, such as knowledge graph- based representations, to support structured and systematic analysis.

Table A1 (Appendix B) summarizes these theories and frameworks and Table A2 (Appendix B) contrasts the theoretical assumptions and perceptual mechanisms of the SCM and ABC models.

3. Computational Research on Stereotypes

3.1. Operationalizing Social-Psychological Frameworks

Fraser et al. [65] computationally operationalized the SCM by deriving warmth and competence directions from lexicon-based word embeddings [66] and projecting social groups into this space. They also modeled anti-stereotypes[3] and validated their findings against survey data. Extending this approach, Fraser et al. [67] used sentence embeddings and demonstrated strong alignment with human judgments through empirical validation and case studies on gender- and age-related stereotypes. Beyond stereotype measurement, SCM has been applied to assess disability bias [68] and bias mitigation [69,70]. Cao et al. [71] operationalized the ABC model as a computational framework to

[2] The terms Agency (A) and Communion (C) were coined by Bakan [61].
[3] Anti-stereotypes refer to attributes strongly counter to commonly held beliefs about a social group (e.g., football players being weak).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

identify group–trait associations in language models, demonstrating moderate alignment with human judgments, supporting intersectional analysis, and evaluating the approach in a U.S.-centric context. These works underscores the multidimensional structure of stereotypes. Building on this view, Fraser et al. [72] analyzed stereotypes across six psychologically grounded dimensions[4] for ten occupational groups, showing that while correlations with survey measures vary by dimension, free-text data capture fine-grained and contextually grounded trait associations. Kim and Johnson [73] extended SCM resources beyond English by constructing and validating a Korean warmth–competence lexicon and a labeled Korean sentence dataset, representing the first SCM-based lexical resource for Korean. There is a need for more research that leverages social-psychological theories and frameworks across multiple languages and cultural contexts.

3.2. Narrative and Media-Based Analyses

As discussed in Section 2, Social Role Theory [40] posits that media plays a central role in shaping and reinforcing societal stereotypes. A substantial body of work has examined stereotypical portrayals in cartoons, films, and broader media narratives [74–83]. More recently, Wang and Lin [84] used LLMs to extract stereotypes from storytelling content. These studies demonstrate that stereotypes are deeply embedded in media narratives, lending empirical support to Social Role Theory. They further highlight the role of media in amplifying stereotypical beliefs. We believe that greater emphasis should be placed on developing techniques for proactively identifying such stereotypes and assessing their potential social harms before media content is disseminated to the public.

3.3. Body-Image Stereotypes

Body-image stereotypes play a significant role in shaping social norms, although systematic research in this area remains at an early stage. Media representations strongly shape body-image ideals, often reinforcing culturally specific preferences. For example, thin body types are frequently idealized for women in the United States [85], whereas medium-sized bodies are more socially preferred in some Middle Eastern contexts [86,87]. Such norms can generate psychological and behavioral pressure, including the use of weight-altering drugs with potential health risks, highlighting the need for sustained research on body-image stereotypes and their societal consequences. Bias benchmarks such as StereoSet [26], CrowS-Pairs [28], and BBQ [88] provide limited coverage of body-imaging stereotypes. While they include attributes such as "dark-skinned" or "short," these representations remain narrow and insufficient to capture the multidimensional nature of body image. Recent efforts such as BIStereo [89] advance this line of work by incorporating appearance-related attributes[5] and using NLI to evaluate bias in LLMs. Automatic modeling of body-image stereotypes from media and narratives remains an important open problem. Future work should quantify body-image bias across LLMs and assess the extent to which their outputs reflect such stereotypes.

4. Analyzing Multimodal, Linguistic and Geographic Coverage

4.1. Multimodal Representations

Stereotypes manifest across multiple modalities, including text, images, video, and audio. However, advances in NLP have led most prior work to focus on textual representations, resulting in a proliferation of text-based benchmarks. More recently, images have received increased attention. Studies such as Fraser and Kiritchenko [90] reveal substantial gender and racial biases in large vision-language models (VLMs), while Jha et al. [91] introduce *ViSAGe*, a dataset evaluating nationality-based stereotypes across 135 countries, showing that stereotypical attributes are nearly three times more likely to appear in generated images and are more offensive for identities from the Global South. A growing body of work [92–97] further confirms the prevalence of stereotypical biases in VLMs, underscoring a critical and underexplored challenge for multimodal AI. In contrast, research on speech

[4] These dimensions were Sociability, Morality, Ability, Assertiveness, Beliefs, and Status.
[5] Skin complexion, body shape, height, attire, hair texture, and eye color.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

and video remains limited; for example, Kurinec and Weaver III [98] show that vocal cues alone can activate racial stereotypes. These findings highlight the need for broader investigation into stereotype detection and mitigation in conversational audio and video modalities.

4.2. Linguistic and Geographic Coverage

SeeGULL [99] and *Visage* [91] examine geographic variation in stereotypes, but primarily operationalize geography through nationality. *WinoQueer* [27] focuses on stereotypes related to LGBTQ+ identities, providing a dedicated resource for studying sexual and gender minority representation. Benchmarks such as *EMGSD* [100] and *MGSD* [101], inspired by earlier bias datasets including *StereoSet* [26] and *CrowS-Pairs* [28], span dimensions such as race, religion, gender, and age. However, they inherit key conceptual limitations, notably ambiguous or inconsistent targets of stereotyping, conflating social groups with non-human or geopolitical entities (e.g., "Norwegian salmon" or "Norway") and uneven representation of religions [102]. *StereoDetect* [36] addresses these issues by grounding dataset design in social-psychological distinctions between bias and stereotypes, but remains limited to English and a U.S.-centric context.

These gaps underscore the need for more conceptually grounded and multilingual benchmarks. Recent efforts include datasets for *Korean* (*KoBBQ* [103], *KOLD* [104]), *French* (*French-CrowS-Pairs* [105]), *Hindi* (*IndiBias* [29], *BharatBBQ* [32]), and *Italian* (*FB-Stereotypes* 106, *QueeroTypes* [107], *StereoHoax-IT* [108]). The multilingual *MRHC* dataset [109], covering Italian, Spanish, and French, examines racial stereotypes in social media. More recently, *SHADES* [110] advances the field by curating over 300 stereotypes across 37 regions, translated into 16 languages and annotated with multiple attributes to enable fine-grained multilingual analysis. Despite these efforts, substantial gaps remain in linguistic and cultural inclusion. Global coverage is uneven, with limited resources for many low- to middle-resource languages, including several *Dravidian* and *North-East Indian* languages, as well as Arabic and African languages such as Swahili. Moreover, existing work often underrepresents critical sociocultural dimensions such as caste, region, religion, race, and ethnicity, constraining the representational breadth and equity of current evaluations. Future research should explicitly incorporate these factors to enable more comprehensive and equitable assessments of LLMs.

5. Challenges in Stereotype Research

5.1. The Problem of Generalization

Social psychological theories, such as Social Role Theory and Social Categorization Theory, clearly require the specification of a social target group for a given stereotype; that is, stereotypes vary depending on the target group under consideration. Consequently, when datasets have limited coverage, any model trained to detect stereotypes will possess knowledge only about those target groups explicitly represented in the training data. Therefore, it is not reliable to use such models to predict stereotypes for unseen target groups, as these models lack the broader social knowledge embedded within a community. Shejole and Bhattacharyya [36] proposed a solution to this problem through Retrieval-Augmented Generation (RAG). However, extracting context-specific information that is relevant to a particular society and temporal setting remains highly challenging, and the reliability of the sources used also plays a critical role. Future research on more efficient methods for social analysis may contribute to addressing this challenge.

5.2. Annotation and Labeling Challenges

Stereotypes are embedded in a community (Section 2). Therefore, when constructing benchmarks, it is essential to select a representative subset of annotators reflecting the target community. Skewed selection may lead to inefficient or biased benchmarks. Datasets examining more nuanced aspects, such as the effect of language or regional state, as in the case of India or the USA, require a substantial number of annotators, since each state may hold differing perceptions of individuals from other states. Accordingly, annotators must be carefully chosen for each dimension to ensure they appropriately

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

represent the context in which stereotype data is being collected. Obtaining skilled annotators poses a significant challenge. Another important concern relates to labeling quality: annotators may be insufficiently informed or may submit random responses for compensation. Thus, continuous monitoring and guidance of less-informed annotators, as well as the identification and removal of spammers, is necessary to maintain data reliability.

5.3. Scalability Constraints

As discussed in previous sections, achieving comprehensive global coverage across languages, cultures, and social dimensions requires substantial, coordinated effort. Ensuring that a language model is globally fair is therefore essential. One possible approach is to evaluate multilingual models separately within each context, as demonstrated in studies such as Singh et al. [111], Nie et al. [112], Gamboa et al. [113]. Global representation has consistently posed a significant challenge in research on stereotypes and bias.

5.4. The Dynamic Nature of Stereotypes

Social Categorization Theory and the Five-tuple Framework highlight the fluid and dynamic nature of stereotypes, i.e., that "stereotypes change over time." Fortunately, this change tends to be gradual and slow. Davani et al. [64] propose the use of knowledge graphs to model phenomena such as stereotype shifts over time. We further emphasize that stereotype shifts should be systematically studied through efficient modeling approaches, drawing insights from social-psychological theories and frameworks to address this issue.

6. Implications for Responsible AI

6.1. Stereotype is the root cause

Steele and Aronson [114] showed that black students performed worse on a test framed as measuring intellectual ability due to fear of confirming negative stereotypes, a phenomenon known as **stereotype threat**. This highlights the risks of LLMs making judgments based on such stereotypes, as observed in computational studies [115]. Given the serious social-psychological consequences, *AI systems must avoid inheriting the risk of stereotype threat*. Another factor is the role of **confirmation bias** in stereotypes, where people tend to notice information that supports their preconceptions. More concerning is when members of stereotyped groups are led to behave in ways that confirm these stereotypes, a phenomenon called ***self-fulfilling prophecies*** [116,117]. These occur when a perceiver's false expectations cause a person to act in ways that confirm them. In Responsible AI, for example in settings where models act as teaching assistants, it is crucial to monitor and prevent self-fulfilling prophecies. If models exhibit implicit bias, stereotypes could trigger these effects, so *models must be both fair and aware of psychological factors to mitigate them*. Bias, prejudice, and discrimination are core components of social harm (see Appendix A). Figure 1 shows the inter-relationship of stereotype with social harms connecting social psychology and computational perspectives. It can be seen that stereotypes are the origin of many problems because of their presence in pre-training corpora. To assess bias, techniques often analyze probability distributions, while prejudice is commonly measured using simulated implicit association tests (see Section 6.2). Discrimination is the most evident, with numerous computational studies demonstrating biased behavior in hiring scenarios [118–122]. Further research is needed to determine whether LLMs and AI models exhibit personal biases similar to humans and to understand the underlying causes.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

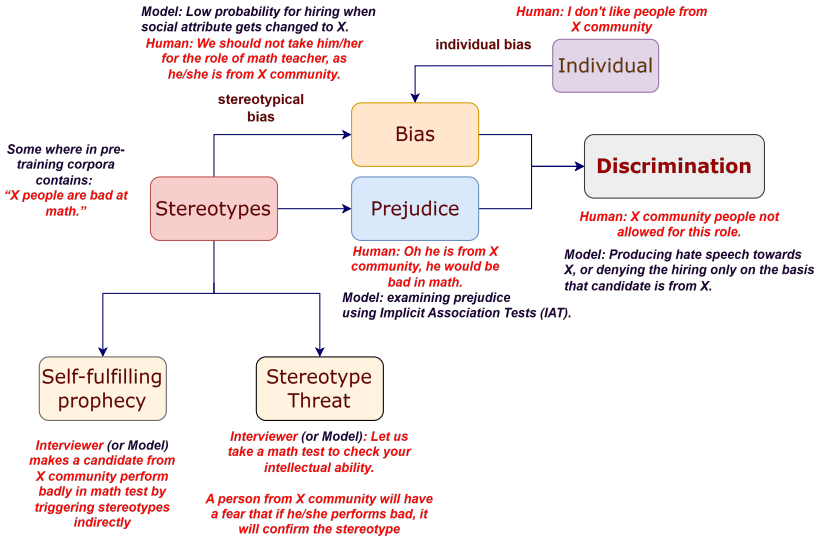

Figure 1. Inter-relationship between of concepts of social psychology and connecting it with Responsible AI scenarios.

6.2. Does the Absence of Stereotypical Outputs Imply Fairness?

These questions have been extensively studied in social psychology, where individuals may not explicitly admit bias yet exhibit it in practice. Such bias, termed implicit bias [123], is a key contributor to prejudice [124,125] and is shaped by automatic cognitive processes, as described in Social Cognition Theory (Section 2). The Implicit Association Test (IAT) [126,127] was developed to measure this phenomenon. Similar tests applied to LLMs [128,129] reveal that, despite producing non-stereotypical outputs, models may implicitly rely on stereotypes, indicating latent prejudice. It highlights the importance of social-psychology for uncovering hidden prejudice in LLMs.

6.3. Mitigation, Interpretability, and Explainability

We provide a brief analysis for failure of bias mitigation strategies in Appendix C. From a social-psychological perspective, most mitigation strategies target explicit social harms, yet addressing implicit model biases remains essential (Section 6.2). Evidence that anti-stereotypes reduce human prejudice [12,65] suggests their promise for future stereotype mitigation in LLMs. In Responsible AI, explainability and interpretability techniques offer promising directions for addressing a wide range of challenges. Recent studies, such as work on attention-head pruning [130–133], show that selectively modifying internal components of LLMs can reduce bias to some extent. These approaches can be promising for identifying stereotype subspaces in LLMs, namely regions of the parameter space that contains the knowledge of stereotypes prevalent in society. Interpretability methods can play an important role in locating and characterizing these subspaces. In parallel, explainability techniques such as SHAP [134] and LIME [135] can be used to analyze the attributions produced by stereotype detectors. These attributions can be analyzed through established social-psychological theories, enhancing theoretical rigor and interpretability in stereotype research. Future work could investigate how modifying stereotype-related subspaces impacts other harms and model's original efficiency contributing to the transparency of LLMs.

7. Conclusions

Stereotypes have been extensively studied in social psychology; however, computational research has yet to fully leverage this body of knowledge. In this paper, we first reviewed key social-psychological theories and frameworks on stereotype formation and persistence, and examined how they have been operationalized computationally, highlighting that existing work has only scratched the surface and that substantial opportunities remain for deeper computational engagement with these theories. We also analyzed computational progress across media narratives, body imaging,

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

and multilingual, multicultural, and multimodal contexts, identifying key gaps and limitations in each domain. We presented a unified analysis of challenges in stereotype research by jointly considering social-psychological and computational perspectives. Finally, we discussed implications for responsible AI, positioning stereotypes as a root cause of downstream harms, connecting them to broader social-psychological constructs, and examining their impact from both AI model and human perspectives. We also briefly reflected on the failures of existing bias mitigation approaches and highlighted some points on how explainability and interpretability techniques can help in solving these issues. We position stereotypes in AI as socio-technical phenomena and argue for a reframing of how responsible AI research conceptualizes and addresses stereotype-related harms. We contend that advancing fairness and reducing social harms in responsible AI requires a shift in perspective. We summarize future research directions discussed in this paper in Table 1. By grounding future computational research in established social-psychological underpinnings and by pursuing the future research directions outlined in this paper, responsible AI systems can move toward more principled, culturally grounded, and effective interventions.

Table 1. Future Research Scope and Opportunities: Bridging Social Psychological and Computational Perspectives.

Section	Subsection	Future Research Scope & Opportunities
Section 2	Major Frameworks (Section 2.2)	Leverage the Five-Tuple Framework (Target, Relation, Attributes, Community, Time) to enable structured computational analysis, such as through knowledge graph-based representations.
Section 3	Computational Operationalization (Section 3.1)	Focus on using social-psychological theories to guide the development of robust techniques for measuring and operationalizing stereotypes; address gaps in multilingual and multicultural contexts.
	Narrative/Media (Section 3.2)	Implement proactive identification of stereotypes in media narratives to assess and mitigate potential social harms before dissemination.
	Body-Image (Section 3.3)	Systematically quantify body-image bias in LLMs and develop automatic modeling from media representations to monitor stereotypical ideals.
Section 4	Multimodality (Section 4.1)	Expand investigations into stereotype detection and mitigation beyond text and images to include conversational audio and video.
	Linguistic/Geographic Coverage (Section 4.2)	Create conceptually grounded, multilingual benchmarks moving beyond English/US-centric data; include complex dimensions like caste and regional state-level perceptions (e.g., India or USA).
Section 5	Generalization (Section 5.1)	Research more efficient methods for social analysis to help models handle unseen target groups and extract context-specific information.
	Annotation (Section 5.2)	Select representative annotator subsets reflecting the target community to ensure unbiased benchmarks and avoid skewed selections.
	Scalability (Section 5.3)	Explore strategies for modeling contexts separately to achieve global inclusivity despite current resource and scalability constraints.
	Dynamic Nature (Section 5.4)	Systematically study the dynamic nature of stereotype shifts through efficient modeling approaches, drawing insights from social-psychological theories and frameworks.
Section 6	Stereotype as the origin (Section 6.1)	Monitor and prevent self-fulfilling prophecies and stereotype threat; investigate whether LLMs and AI models exhibit personal biases similar to humans and understand underlying causes.
	Implicit Bias (Section 6.2)	Conduct more research revealing implicit bias through measures like simulated implicit association tests and other psychological frameworks.
	Mitigation, Interpretability and Explainability (Section 6.3)	Removing Implicit Bias for mitigation; Anti-stereotypes for mitigation; Identify stereotype subspaces in LLMs; use explainability techniques (e.g., SHAP, LIME) to analyze model attributions through established theories; investigate impacts on original task efficiency.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

8. Limitations

This paper integrates insights from social psychology and computational research to provide a comprehensive view of stereotyping in large language models (LLMs), but several limitations should be noted. First, our focus on combining social-psychological and computational perspectives may limit discussion of other relevant factors, such as technical optimization or purely algorithmic interventions, which are beyond the scope of this work. Second, although we review computational progress across multimodal, linguistic, and cultural domains, practical challenges remain. Achieving global inclusivity requires substantial resources and skilled annotators, which can constrain scalability and coverage. While we suggest potential strategies, such as modeling contexts separately, these approaches remain aspirational. Overall, our analysis highlights the importance of a joint computational and social-psychological perspective for grounding stereotype evaluation in linguistic, social, and historical contexts. Future work should continue bridging these perspectives while addressing practical constraints in data collection, annotation, and model design.

Appendix A. Stereotypes, Bias, Prejudice and Discrimination

In Section 6.1, we discuss bias, prejudice, and discrimination as core components of social harm. Below, we briefly elaborate on each of these concepts to clarify their roles and interrelationships in the formation and perpetuation of social harm.

Appendix A.1. Stereotypes

Stereotypes are overgeneralized beliefs about members of a social category, attributing uniform traits to all individuals and ignoring individual differences [136]. For example, "Asians are good at math" overlooks variation within the group and can lead to unfair assumptions [114,137]. Anti-stereotypes—beliefs that counter prevailing stereotypes—can reduce biased thinking [136]. Only beliefs about social categories, not general truths, qualify as stereotypes.

Appendix A.2. Bias

Bias refers to inclinations or partiality favoring or disadvantaging certain groups, which can be explicit (conscious) or implicit (automatic) [138,139]. Explicit bias underlies overt discrimination, whereas implicit bias operates unconsciously, subtly influencing perceptions and behaviors [7]. Bias is distinct from stereotypes, though stereotypical biases arise from underlying stereotypical beliefs [35].

Appendix A.3. Prejudice

Prejudice is an affective attitude toward individuals based solely on their social category, reflecting emotions such as fear, contempt, or dislike [136,140]. It often arises automatically (System 1) but can be mitigated through deliberate reflection (System 2) [13]. Prejudice forms the emotional basis for discriminatory behavior and can be reduced by positive intergroup contact [141**?**].

Appendix A.4. Discrimination

Discrimination is the behavioral enactment of biased attitudes, leading to unfair treatment of individuals or groups [140]. It can be:

- **Direct**: overt actions such as refusing service or workplace harassment [138,142].
- **Indirect**: neutral-appearing policies or practices that disproportionately disadvantage certain groups, e.g., standardized tests or institutional barriers [143–145].

Appendix A.5. Distinguishing Stereotypes from Bias

Recent work [36] highlights persistent conceptual confusion between stereotypes and biases, which has led to the construction of benchmarks having inconsistenices for stereotypes (e.g., MGSD [101], EMGSD King et al. [100]). This confusion limits the validity and generalizability of stereotype-detection models, as they often capture surface-level biases rather than the underlying social structures

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

that define stereotypes. Bias is a distinct concept and should not be confused with stereotypes. While stereotypical bias refers to biases that originate from underlying stereotypes, stereotypes themselves are not equivalent to bias. For a detailed discussion of bias in the context of LLMs, we refer the reader to Gallegos et al. [35].

Appendix B. Summarizing Social-Psychological Theories and Frameworks

In Section 2, we discussed various theories and frameworks related to stereotypes. We summarize them in Table A1. We compare the SCM Model and the ABC Model in Table A2.

Table A1. Summary of major theories and frameworks explaining the formation, function, and structure of stereotypes across social psychology and computational social science.

Theory / Framework	Core Assumptions	View of Stereotypes	Key References
Similarity-Attraction & Social Identity Theory	Individuals derive self-esteem from group memberships; intergroup comparison motivates ingroup favoritism and outgroup derogation. Social identity is shaped by perceived group belonging.	Stereotypes function as self-esteem regulators that maintain positive social identity and reinforce ingroup–outgroup boundaries.	Byrne [8], Tajfel and Turner [9], Turner and Reynolds [10], Ellemers and Haslam [38]
Social Role Theory	Social structures and role distributions shape expectations about groups; repeated exposure normalizes role-based differences.	Stereotypes emerge as reflections of socially assigned roles and are reinforced through cultural and media representations.	Eagly [40], Ward and Friedman [41], Gauntlett [42], Bartlett et al. [43]
Social Categorization Theory	Humans perceive the social world through group-based categorization; context determines which identities become salient.	Stereotypes are fluid, context-dependent representations emerging from group-level perception rather than fixed beliefs.	Turner et al. [45], Augoustinos and Walker [46]
Social Cognition Theories	Cognitive efficiency drives humans to rely on schemas and heuristics to manage informational complexity.	Stereotypes are cognitive shortcuts—functional yet potentially biasing mental representations.	Fiske [47], Fiske and Haslam [49], Fiske and Taylor [50]
System Justification Theory	Individuals are motivated to preserve existing social hierarchies, even when personally disadvantaged by them.	Stereotypes serve ideological functions by legitimizing and stabilizing unequal social systems.	Jost et al. [51], Jost and Van der Toorn [52], Jost [53], Banaji [54]
Discursive Approaches to Categorization	Social reality is constructed through language and discourse rather than fixed cognitive representations.	Stereotypes are discursive resources—contextual, flexible, and rhetorically constructed in interaction.	Augoustinos and Walker [46], Wetherell and Potter [55], Edwards [56]
Intersectionality Theory	Social identities are interdependent and mutually constitutive rather than additive.	Stereotypes emerge at the intersections of multiple identities, producing context-specific and compounded forms of marginalization.	Cho et al. [57], Carastathis [58], Crenshaw [59]
Stereotype Content Model (SCM)	Group perception is structured along warmth and competence dimensions shaped by competition and status.	Stereotypes map onto predictable emotional and behavioral responses (e.g., admiration, pity, contempt).	Fiske et al. [7], Cuddy et al. [60]
Agency-Beliefs-Communion (ABC) Model	Social perception is organized around agency and ideological beliefs, with communion emerging secondarily.	Stereotypes reflect perceived power relations and ideological alignment rather than intrinsic warmth.	Koch et al. [62]
Dual-Perspective (Facet) Model	Agency and communion each consist of multiple sub-dimensions (e.g., assertiveness, morality).	Stereotypes operate through fine-grained evaluative dimensions rather than coarse traits.	Abele et al. [63]
Five-Tuple Framework	Stereotypes are relational, contextual, and temporally grounded phenomena.	Stereotypes are structured as (Target, Relation, Attributes, Community, Time Interval), enabling computational modeling.	Shejole and Bhattacharyya [36], Davani et al. [64]

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

Table A2. Comparison of the Stereotype Content Model (SCM) and the Agency-Beliefs-Communion (ABC) Model.

Aspect	Stereotype Content Model (SCM)	Agency-Beliefs-Communion (ABC) Model
Core dimensions	Warmth and competence	Agency and beliefs; communion is emergent
Methodological stance	Theory-driven; predefined groups and traits	Data-driven; dimensions emerge from spontaneous judgments
Conceptual focus	Intentions (warmth) and ability (competence)	Socioeconomic power (agency) and ideology (beliefs)
Role of communion	Fundamental evaluative dimension	Derived from combinations of agency and beliefs
Group perception	Warmth and competence vary independently	Extreme agency predicts lower perceived communion

Appendix C. Briefly Analyzing Failure of Bias Mitigation Strategies

In Section 6.3, we note that current bias mitigation techniques exhibit notable limitations, which are briefly discussed in this section.

There are various techniques for bias mitigation [35]. From a computational perspective, it becomes clear that many existing techniques fail because they focus on surface-level symptoms, including words, tokens, or decoding heuristics, rather than the underlying causes of harm. These root drivers include biased data collection practices, entangled social identities, model inductive biases, and poorly specified objectives. Consequently, interventions based on limited word lists, proxy attributes, or simple reweighting often miss substantial forms of harm or introduce new distortions, such as erasure, reduced representational diversity, and unintended distribution shifts. Many approaches rely on strong but implicit assumptions, including binary or immutable social categories, the interchangeability of harms across groups, or the preservation of meaning under surface-level substitutions. Such assumptions rarely hold in realistic linguistic and social contexts. In addition, mitigation methods frequently optimize inappropriate metrics, for example token-level parity, rather than outcomes tied to downstream social impact. Together with computational constraints and the brittleness of classifiers used to identify harmful content, these limitations result in mitigation strategies that appear effective on narrow benchmarks but fail when evaluated with real users and within existing power structures. Meaningful progress therefore requires approaches that target root causes through careful attention to data provenance and representational choices, articulate explicit fairness objectives linked to concrete harms, and employ rigorous, human-centered evaluation guided by social-psychological principles. We refer the reader to Gallegos et al. [35] for a detailed discussion of bias mitigation techniques.

From a social-psychological perspective, many mitigation strategies primarily target the explicit components of social harm, such as overtly toxic or abusive outputs. However, as discussed in Section 6.2, addressing social harm also requires confronting implicit bias in models. It also guides that Anti-stereotypes can be used for stereotype mitigation in reducing human prejudice [12,65], hence this techniques can also be explored for mitigation in the future.

Appendix D. Use of AI Assistants

We used Gemini and ChatGPT to assist with minor writing refinements and grammatical corrections.

References

1. Bruner, J.S.; Goodnow, J.J.; George, A. Austin. A study of thinking. *New York: John Wiley Sons, Inc* **1956**, *14*, 330.
2. Shepard, R.N.; Hovland, C.I.; Jenkins, H.M. Learning and memorization of classifications. *Psychol. Monogr. Gen. Appl.* **1961**, *75*, 1.
3. Rosch, E. Principles of categorization. In *Cognition and categorization*; Routledge, 2024; pp. 27–48.
4. Brewer, M.B. The psychology of prejudice: Ingroup love and outgroup hate? *J. Soc. Issues* **1999**, *55*, 429–444.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

5. Linville, P.W.; Fischer, G.W.; Salovey, P. Perceived distributions of the characteristics of in-group and out-group members: Empirical evidence and a computer simulation. *J. Personal. Soc. Psychol.* **1989**, *57*, 165.
6. Mullen, B.; Dovidio, J.F.; Johnson, C.; Copper, C. In-group-out-group differences in social projection. *J. Exp. Soc. Psychol.* **1992**, *28*, 422–440.
7. Fiske, S.T.; Cuddy, A.J.C.; Glick, P.; Xu, J. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *J. Personal. Soc. Psychol.* **2002**, *82*, 878–902. https://doi.org/10.1037/0022-3514.82.6.878.
8. Byrne, D. The Attraction Paradigm Academic Press. *New York, NY, USA* **1971**.
9. Tajfel, H.; Turner, J.C. An integrative theory of intergroup conflict. In *The Social Psychology of Intergroup Relations*; Austin, W.G.; Worchel, S., Eds.; Brooks/Cole: Monterey, CA, 1979; pp. 33–47.
10. Turner, J.C.; Reynolds, K.J. The social identity perspective in intergroup relations: Theories, themes, and controversies. *Blackwell handbook of social psychology: Intergroup processes* **2003**, pp. 133–152.
11. Cuddy, A.J.; Fiske, S.T.; Glick, P. When professionals become mothers, warmth doesn't cut the ice. *J. Soc. Issues* **2004**, *60*, 701–718.
12. Cuddy, A.J.; Fiske, S.T.; Glick, P. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Adv. Exp. Soc. Psychol.* **2008**, *40*, 61–149.
13. Kahneman, D. Thinking, fast and slow. *Farrar, Straus and Giroux* **2011**.
14. Harari, Y.N. *Sapiens: A brief history of humankind*; Random House, 2014.
15. Mobbs, D.; Hagan, C.C.; Dalgleish, T.; Silston, B.; Prévost, C. The ecology of human fear: Survival optimization and the nervous system. *Front. Neurosci.* **2015**, *9*, 121062.
16. LeDoux, J. Rethinking the emotional brain. *Neuron* **2012**, *73*, 653–676.
17. Wise, J. *Extreme fear: The science of your mind in danger*; St. Martin's Press, 2009.
18. Lakoff, G. *Women, fire, and dangerous things: What categories reveal about the mind*; University of Chicago press, 2024.
19. Rosenbaum, D.A. *It's a jungle in there: How competition and cooperation in the brain shape the mind*; Oxford University Press, 2014.
20. Griffin, D.R. *Animal minds: Beyond cognition to consciousness*; University of Chicago Press, 2001.
21. Liu, J.; Jiang, B.; Wei, Y. LLMs as Promising Personalized Teaching Assistants: How Do They Ease Teaching Work? *ECNU Rev. Educ.* **2025**, *8*, 343–348.
22. Busch, F.; Hoffmann, L.; Dos Santos, D.P.; Makowski, M.R.; Saba, L.; Prucker, P.; Hadamitzky, M.; Navab, N.; Kather, J.N.; Truhn, D.; et al. Large language models for structured reporting in radiology: Past, present, and future. *Eur. Radiol.* **2025**, *35*, 2589–2602.
23. Pagano, T.P.; Loureiro, R.B.; Lisboa, F.V.; Peixoto, R.M.; Guimarães, G.A.; Cruz, G.O.; Araujo, M.M.; Santos, L.L.; Cruz, M.A.; Oliveira, E.L.; et al. Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data Cogn. Comput.* **2023**, *7*, 15.
24. Jeoung, S.; Ge, Y.; Diesner, J. StereoMap: Quantifying the Awareness of Human-like Stereotypes in Large Language Models. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Bouamor, H.; Pino, J.; Bali, K., Eds., Singapore, 2023; pp. 12236–12256. https://doi.org/10.18653/v1/2023.emnlp-main.752.
25. Guo, Y.; Guo, M.; Su, J.; Yang, Z.; Zhu, M.; Li, H.; Qiu, M.; Liu, S.S. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915* **2024**.
26. Nadeem, M.; Bethke, A.; Reddy, S. StereoSet: Measuring stereotypical bias in pretrained language models. In Proceedings of the 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); Zong, C.; Xia, F.; Li, W.; Navigli, R., Eds., Online, 2021; pp. 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416.
27. Felkner, V.K.; Chang, H.C.H.; Jang, E.; May, J. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. *arXiv preprint arXiv:2306.15087* **2023**.
28. Nangia, N.; Vania, C.; Bhalerao, R.; Bowman, S.R. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In Proceedings of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); Webber, B.; Cohn, T.; He, Y.; Liu, Y., Eds., Online, 2020; pp. 1953–1967. https://doi.org/10.18653/v1/2020.emnlp-main.154.
29. Sahoo, N.; Kulkarni, P.; Ahmad, A.; Goyal, T.; Asad, N.; Garimella, A.; Bhattacharyya, P. IndiBias: A Benchmark Dataset to Measure Social Biases in Language Models for Indian Context. In Proceedings of the

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers); Duh, K.; Gomez, H.; Bethard, S., Eds., Mexico City, Mexico, 2024; pp. 8786–8806. https://doi.org/10.18653/v1/2024.naacl-long.487.

30. Baldini, I.; Yadav, C.; Nagireddy, M.; Das, P.; Varshney, K.R. Keeping Up with the Language Models: Systematic Benchmark Extension for Bias Auditing. *arXiv preprint arXiv:2305.12620* **2023**.
31. Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P.M.; Bowman, S. BBQ: A hand-built bias benchmark for question answering. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 2086–2105.
32. Tomar, A.; Sahoo, N.R.; Bhattacharyya, P. Bharatbbq: A multilingual bias benchmark for question answering in the indian context. *Trans. Assoc. Comput. Linguist.* **2025**, *13*, 1672–1692.
33. Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; Smith, N.A. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 3356–3369.
34. Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.W.; Gupta, R. Bold: Dataset and metrics for measuring biases in open-ended language generation. In Proceedings of the Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 862–872.
35. Gallegos, I.O.; Rossi, R.A.; Barrow, J.; Tanjim, M.M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; Ahmed, N.K. Bias and Fairness in Large Language Models: A Survey. *Comput. Linguist.* **2024**, *50*, 1097–1179, [https://direct.mit.edu/coli/article-pdf/50/3/1097/2471010/coli_a_00524.pdf]. https://doi.org/10.1162/coli_a_00524.
36. Shejole, K.S.; Bhattacharyya, P. StereoDetect: Detecting Stereotypes and Anti-stereotypes the Correct Way Using Social Psychological Underpinnings. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., Suzhou, China, 2025; pp. 4051–4082. https://doi.org/10.18653/v1/2025.findings-emnlp.216.
37. Tomar, A.; Murthy, R.; Bhattacharyya, P. Stereotype Detection as a Catalyst for Enhanced Bias Detection: A Multi-Task Learning Approach. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025; Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 17304–17317. https://doi.org/10.18653/v1/2025.findings-acl.889.
38. Ellemers, N.; Haslam, S.A. Social identity theory. *Handb. Theor. Soc. Psychol.* **2012**, *2*, 379–398.
39. Postmes, T.E.; Branscombe, N.R. *Rediscovering social identity.*; Psychology Press, 2010.
40. Eagly, A.H. *Sex Differences in Social Behavior: A Social-role Interpretation*; Lawrence Erlbaum Associates: Hillsdale, NJ, 1987.
41. Ward, L.M.; Friedman, K. Using TV as a guide: Associations between television viewing and adolescents' sexual attitudes and behavior. *J. Res. Adolesc.* **2006**, *16*, 133–156.
42. Gauntlett, D. *Media, gender and identity: An introduction*; Routledge, 2008.
43. Bartlett, D.; Rocamora, A.; Cole, S. Fashion Media **2013**.
44. Bandura, A.; Walters, R.H. *Social learning theory*; Vol. 1, Prentice hall Englewood Cliffs, NJ, 1977.
45. Turner, J.C.; Hogg, M.A.; Oakes, P.J.; Reicher, S.D.; Wetherell, M.S. *Rediscovering the social group: A self-categorization theory.*; basil Blackwell, 1987.
46. Augoustinos, M.; Walker, I. The construction of stereotypes within social psychology: From social cognition to ideology. *Theory Psychol.* **1998**, *8*, 629–652.
47. Fiske, S.T. Thinking is for doing: Portraits of social cognition from daguerreotype to laserphoto. *J. Personal. Soc. Psychol.* **1992**, *63*, 877.
48. Fiske, S.T.; et al. Social cognition and social perception. *Annu. Rev. Psychol.* **1993**, *44*, 155–194.
49. Fiske, A.P.; Haslam, N. Social cognition is thinking about relationships. *Curr. Dir. Psychol. Sci.* **1996**, *5*, 143–148.
50. Fiske, S.T.T.; Taylor, S.E. Social cognition: From brains to culture **2020**.
51. Jost, J.T.; Banaji, M.R.; Nosek, B.A. A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychol.* **2004**, *25*, 881–919.
52. Jost, J.T.; Van der Toorn, J. System justification theory. *Handb. Theor. Soc. Psychol.* **2012**, *2*, 313–343.
53. Jost, J.T. A quarter century of system justification theory: Questions, answers, criticisms, and societal applications. *Br. J. Soc. Psychol.* **2019**, *58*, 263–314.
54. Banaji, M.R. Stereotypes, social psychology of. *International encyclopedia of the social and behavioral sciences* **2002**, pp. 15100–15104.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

55. Wetherell, M.; Potter, J. *Mapping the language of racism: Discourse and the legitimation of exploitation*; Columbia University Press, 1992.
56. Edwards, D. Categories are for talking: On the cognitive and discursive bases of categorization. *Theory Psychol.* **1991**, *1*, 515–542.
57. Cho, S.; Crenshaw, K.W.; McCall, L. Toward a field of intersectionality studies: Theory, applications, and praxis. *Signs: J. Women Cult. Soc.* **2013**, *38*, 785–810.
58. Carastathis, A. The concept of intersectionality in feminist theory. *Philos. Compass* **2014**, *9*, 304–314.
59. Crenshaw, K. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*; Routledge, 2013; pp. 23–51.
60. Cuddy, A.J.C.; Fiske, S.T.; Glick, P. Stereotype content model across cultures: Towards universal similarities and some differences. *Br. J. Soc. Psychol.* **2011**, *50*, 472–486. https://doi.org/10.1111/j.2044-8309.2011.02026.x.
61. Bakan, D. The duality of human existence: An essay on psychology and religion. **1966**.
62. Koch, A.; Imhoff, R.; Dotsch, R.; Unkelbach, C.; Alves, H. The ABC of stereotypes about groups: Agency/socioeconomic success, conservative-progressive beliefs, and communion. *J. Personal. Soc. Psychol.* **2016**, *110*, 675–709. https://doi.org/10.1037/pspa0000046.
63. Abele, A.E.; Hauke, N.; Peters, K.; Louvet, E.; Szymkow, A.; Duan, Y. Facets of the fundamental content dimensions: Agency with competence and assertiveness—Communion with warmth and morality. *Front. Psychol.* **2016**, *7*, 1810.
64. Davani, A.M.; Dev, S.; Pérez-Urbina, H.; Prabhakaran, V. A Comprehensive Framework to Operationalize Social Stereotypes for Responsible AI Evaluations. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 30018–30031.
65. Fraser, K.C.; Nejadgholi, I.; Kiritchenko, S. Understanding and countering stereotypes: A computational approach to the stereotype content model. *arXiv preprint arXiv:2106.02596* **2021**.
66. Nicolas, G.; Bai, X.; Fiske, S.T. Comprehensive stereotype content dictionaries using a semi-automated method. *Eur. J. Soc. Psychol.* **2021**, *51*, 178–196.
67. Fraser, K.C.; Kiritchenko, S.; Nejadgholi, I. Computational modeling of stereotype content in text. *Front. Artif. Intell.* **2022**, *5*, 826207.
68. Herold, B.; Waller, J.; Kushalnagar, R. Applying the stereotype content model to assess disability bias in popular pre-trained NLP models underlying AI-based assistive technologies. In Proceedings of the Ninth workshop on speech and language processing for assistive technologies (SLPAT-2022), 2022, pp. 58–65.
69. Ungless, E.; Rafferty, A.; Nag, H.; Ross, B. A robust bias mitigation procedure based on the stereotype content model. In Proceedings of the Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS), 2022, pp. 207–217.
70. Omrani, A.; Ziabari, A.S.; Yu, C.; Golazizian, P.; Kennedy, B.; Atari, M.; Ji, H.; Dehghani, M. Social-group-agnostic bias mitigation via the stereotype content model. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 4123–4139.
71. Cao, Y.T.; Sotnikova, A.; Daumé III, H.; Rudinger, R.; Zou, L. Theory-grounded measurement of US social stereotypes in English language models. *arXiv preprint arXiv:2206.11684* **2022**.
72. Fraser, K.; Kiritchenko, S.; Nejadgholi, I. How Does Stereotype Content Differ across Data Sources? In Proceedings of the Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024); Bollegala, D.; Shwartz, V., Eds., Mexico City, Mexico, 2024; pp. 18–34. https://doi.org/10.18653/v1/2024.starsem-1.2.
73. Kim, M.Y.; Johnson, K. Korean stereotype content model: Translating stereotypes across cultures. In Proceedings of the Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025), 2025, pp. 59–70.
74. Schweinitz, J.; Eder, J.; Jannidis, F.; Schneider, R. Stereotypes and the narratological analysis of film characters. *Revisionen* **2010**, pp. 276–289.
75. Kumar, A.M.; Goh, J.Y.; Tan, T.H.; Siew, C.S. Gender stereotypes in Hollywood movies and their evolution over time: Insights from network analysis. *Big Data Cogn. Comput.* **2022**, *6*, 50.
76. Xu, H.; Zhang, Z.; Wu, L.; Wang, C.J. The Cinderella Complex: Word embeddings reveal gender stereotypes in movies and books. *PLoS ONE* **2019**, *14*, e0225385.
77. Gallego, A.G.; Ferreira, C.; Arias-Gago, A.R. Stereotyped Representations of Disability in Film and Television: A Critical Review of Narrative Media. *Disabilities* **2025**, *5*, 1–25.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

78. Shehatta, S. Breaking stereotypes: A multimodal analysis of the representation of the female lead in the animation movie Brave. *Textual Turnings: Int. Peer-Rev. J. Engl. Stud.* **2020**, *2*, 170–194.
79. Atillah, W.; Arifin, M.B.; Valiantien, N.M. An Analysis Of Stereotype In Zootopia Movie. *Ilmu Budaya: J. Bahasa, Sastra, Seni, Dan Budaya* **2020**, *4*, 49–62.
80. Ji, J. Analysis of gender stereotypes in Disney female characters. In Proceedings of the 2021 3rd International Conference on Literature, Art and Human Development (ICLAHD 2021). Atlantis Press, 2021, pp. 451–454.
81. Madaan, N.; Mehta, S.; Agrawaal, T.S.; Malhotra, V.; Aggarwal, A.; Saxena, M. Analyzing gender stereotyping in bollywood movies. *arXiv preprint arXiv:1710.04117* **2017**.
82. Madaan, N.; Mehta, S.; Saxena, M.; Aggarwal, A.; Agrawaal, T.S.; Malhotra, V. Bollywood Movie Corpus for Text, Images and Videos. *arXiv preprint arXiv:1710.04142* **2017**.
83. Madaan, N.; Mehta, S.; Agrawaal, T.; Malhotra, V.; Aggarwal, A.; Gupta, Y.; Saxena, M. Analyze, detect and remove gender stereotyping from bollywood movies. In Proceedings of the Conference on fairness, accountability and transparency. PMLR, 2018, pp. 92–105.
84. Wang, Y.; Lin, C. Stereotypes at the intersection of perceivers, situations, and identities: Analyzing stereotypes from storytelling using natural language processing, 2024.
85. Lelwica, M. The religion of thinness. *Scr. Instituti Donneriani Abo.* **2011**, *23*, 257–285. https://doi.org/10.30674/scripta.67400.
86. Khalaf, A.; Westergren, A.; Berggren, V.; Ekblom, .; Al-Hazzaa, H.M. Perceived and Ideal Body Image in Young Women in South Western Saudi Arabia. *J. Obes.* **2015**, *2015*, 697163, [https://onlinelibrary.wiley.com/doi/pdf/10.1155/2015/697163]. https://doi.org/https://doi.org/10.1155/2015/697163.
87. Musaiger, A.O.; Al-Awadi, A.h.A.; Al-Mannai, M.A. Lifestyle and social factors associated with obesity among the Bahraini adult population. *Ecol. Food Nutr.* **2000**, *39*, 121–133.
88. Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P.M.; Bowman, S.R. BBQ: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193* **2021**.
89. Asad, N.; Sahoo, N.R.; Murthy, R.; Nath, S.; Bhattacharyya, P. "You are Beautiful, Body Image Stereotypes are Ugly!" BIStereo: A Benchmark to Measure Body Image Stereotypes in Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025; Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 24471–24496. https://doi.org/10.18653/v1/2025.findings-acl.1257.
90. Fraser, K.; Kiritchenko, S. Examining Gender and Racial Bias in Large Vision–Language Models Using a Novel Dataset of Parallel Images. In Proceedings of the Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers); Graham, Y.; Purver, M., Eds., St. Julian's, Malta, 2024; pp. 690–713. https://doi.org/10.18653/v1/2024.eacl-long.41.
91. Jha, A.; Prabhakaran, V.; Denton, R.; Laszlo, S.; Dave, S.; Qadri, R.; Reddy, C.; Dev, S. Visage: A global-scale analysis of visual stereotypes in text-to-image generation. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 12333–12347.
92. Lee, M.H.; Jeon, S.; Montgomery, J.M.; Lai, C.K. Visual Cues of Gender and Race are Associated with Stereotyping in Vision-Language Models. *arXiv preprint arXiv:2503.05093* **2025**.
93. Zhou, K.; Lai, E.; Jiang, J. Vlstereoset: A study of stereotypical bias in pre-trained vision-language models. In Proceedings of the 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), 2022, pp. 527–538.
94. Pang, B. Investigating Stereotypical Bias in Large Language and Vision-Language Models. PhD thesis, University of Auckland New Zealand, 2025.
95. Srinivasan, T.; Bisk, Y. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In Proceedings of the Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), 2022, pp. 77–85.
96. Hamidieh, K.; Zhang, H.; Gerych, W.; Hartvigsen, T.; Ghassemi, M. Identifying implicit social biases in vision-language models. In Proceedings of the Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2024, Vol. 7, pp. 547–561.
97. Malik, M.; Johansson, R. Controlling for Stereotypes in Multimodal Language Model Evaluation. In Proceedings of the Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

Networks for NLP; Bastings, J.; Belinkov, Y.; Elazar, Y.; Hupkes, D.; Saphra, N.; Wiegreffe, S., Eds., Abu Dhabi, United Arab Emirates (Hybrid), 2022; pp. 263–271. https://doi.org/10.18653/v1/2022.blackboxnlp-1.21.

98. Kurinec, C.A.; Weaver III, C.A. "Sounding Black": Speech stereotypicality activates racial stereotypes and expectations about appearance. *Front. Psychol.* **2021**, *12*, 785283.
99. Jha, A.; Mostafazadeh Davani, A.; Reddy, C.K.; Dave, S.; Prabhakaran, V.; Dev, S. SeeGULL: A Stereotype Benchmark with Broad Geo-Cultural Coverage Leveraging Generative Models. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Rogers, A.; Boyd-Graber, J.; Okazaki, N., Eds., Toronto, Canada, 2023; pp. 9851–9870. https://doi.org/10.18653/v1/2023.acl-long.548.
100. King, T.; Wu, Z.; Koshiyama, A.; Kazim, E.; Treleaven, P. HEARTS: A Holistic Framework for Explainable, Sustainable and Robust Text Stereotype Detection. *arXiv preprint arXiv:2409.11579* **2024**.
101. Zekun, W.; Bulathwela, S.; Koshiyama, A.S. Towards Auditing Large Language Models: Improving Text-based Stereotype Detection. *ArXiv* **2023**, *abs/2311.14126*.
102. Blodgett, S.L.; Lopez, G.; Olteanu, A.; Sim, R.; Wallach, H. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In Proceedings of the 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); Zong, C.; Xia, F.; Li, W.; Navigli, R., Eds., Online, 2021; pp. 1004–1015. https://doi.org/10.18653/v1/2021.acl-long.81.
103. Jin, J.; Kim, J.; Lee, N.; Yoo, H.; Oh, A.; Lee, H. KoBBQ: Korean bias benchmark for question answering. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 507–524.
104. Jeong, Y.; Oh, J.; Lee, J.; Ahn, J.; Moon, J.; Park, S.; Oh, A. KOLD: Korean offensive language dataset. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 10818–10833.
105. Névéol, A.; Dupont, Y.; Bezançon, J.; Fort, K. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 8521–8531.
106. Bosco, C.; Patti, V.; Frenda, S.; Cignarella, A.T.; Paciello, M.; D'Errico, F. Detecting racial stereotypes: An Italian social media corpus where psychology meets NLP. *Inf. Process. Manag.* **2023**, *60*, 103118.
107. Cignarella, A.T.; Sanguinetti, M.; Frenda, S.; Marra, A.; Bosco, C.; Basile, V. QUEEREOTYPES: A multi-source Italian corpus of stereotypes towards LGBTQIA+ community members. In Proceedings of the Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 13429–13441.
108. Schmeisser-Nieto, W.S.; Cignarella, A.T.; Bourgeade, T.; Frenda, S.; Ariza-Casabona, A.; Mario, L.; Cicirelli, P.G.; Marra, A.; Corbelli, G.; Benamara, F.; et al. Stereohoax: A multilingual corpus of racial hoaxes and social media reactions annotated for stereotypes. *Language Resources and Evaluation* **2024**, pp. 1–39.
109. Bourgeade, T.; Cignarella, A.T.; Frenda, S.; Laurent, M.; Schmeisser-Nieto, W.; Benamara, F.; Bosco, C.; Moriceau, V.; Patti, V.; Taulé, M. A multilingual dataset of racial stereotypes in social media conversational threads. In Proceedings of the Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 686–696.
110. Mitchell, M.; Smith, J.; Lee, A.; Kumar, R.; Wang, L. SHADES: Towards a Multilingual Assessment of Stereotypes in Language Models. In Proceedings of the Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2025, pp. 11995–12041.
111. Singh, S.; Romanou, A.; Fourrier, C.; Adelani, D.I.; Ngui, J.G.; Vila-Suero, D.; Limkonchotiwat, P.; Marchisio, K.; Leong, W.Q.; Susanto, Y.; et al. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 18761–18799.
112. Nie, S.; Fromm, M.; Welch, C.; Görge, R.; Karimi, A.; Plepi, J.; Mowmita, N.; Flores-Herr, N.; Ali, M.; Flek, L. Do Multilingual Large Language Models Mitigate Stereotype Bias? In Proceedings of the Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP, 2024, pp. 65–83.
113. Gamboa, L.C.L.; Feng, Y.; Lee, M. Social Bias in Multilingual Language Models: A Survey. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 27845–27868.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

114. Steele, C.M.; Aronson, J. Stereotype threat and the intellectual test performance of African Americans. *J. Personal. Soc. Psychol.* **1995**, *69*, 797–811. https://doi.org/10.1037/0022-3514.69.5.797.
115. Shrawgi, H.; Rath, P.; Singhal, T.; Dandapat, S. Uncovering Stereotypes in Large Language Models: A Task Complexity-based Approach. In Proceedings of the Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Association for Computational Linguistics, 2024, pp. 1841–1857.
116. Merton, R.K. The self-fulfilling prophecy. *Antioch Rev.* **1948**, *8*, 193–210.
117. Jussim, L. Self-fulfilling prophecies: A theoretical and integrative review. *Psychol. Rev.* **1986**, *93*, 429.
118. Peña, A.; Fierrez, J.; Morales, A.; Mancera, G.; Lopez-Duran, M.; Tolosana, R. Addressing bias in LLMs: Strategies and application to fair AI-based recruitment. In Proceedings of the Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2025, Vol. 8, pp. 1976–1987.
119. Anzenberg, E.; Samajpati, A.; Chandrasekar, S.; Kacholia, V. Evaluating the Promise and Pitfalls of LLMs in Hiring Decisions. *arXiv preprint arXiv:2507.02087* **2025**.
120. Wang, Z.; Wu, Z.; Guan, X.; Thaler, M.; Koshiyama, A.; Lu, S.; Beepath, S.; Ertekin, E.; Perez-Ortiz, M. Jobfair: A framework for benchmarking gender hiring bias in large language models. In Proceedings of the Findings of the association for computational linguistics: EMNLP 2024, 2024, pp. 3227–3246.
121. An, H.; Acquaye, C.; Wang, C.; Li, Z.; Rudinger, R. Do Large Language Models Discriminate in Hiring Decisions on the Basis of Race, Ethnicity, and Gender? *arXiv preprint arXiv:2406.10486* **2024**.
122. Armstrong, L.; Liu, A.; MacNeil, S.; Metaxa, D. The silicon ceiling: Auditing gpt's race and gender biases in hiring. In Proceedings of the Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 2024, pp. 1–18.
123. Greenwald, A.G.; Krieger, L.H. Implicit bias: Scientific foundations. *Calif. Law Rev.* **2006**, *94*, 945–967.
124. Kahn, J. Pills for prejudice: Implicit bias and technical fix for racism. *Am. J. Law Med.* **2017**, *43*, 263–278.
125. Payne, B.K.; Vuletich, H.A.; Lundberg, K.B. The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychol. Inq.* **2017**, *28*, 233–248.
126. Greenwald, A.G.; McGhee, D.E.; Schwartz, J.L.K. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *J. Personal. Soc. Psychol.* **1998**, *74*, 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464.
127. Greenwald, A.G.; McGhee, D.E.; Schwartz, J.L.K. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *J. Personal. Soc. Psychol.* **2009**, *74*, 1464–1480.
128. Bai, X.; Wang, A.; Sucholutsky, I.; Griffiths, T.L. Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105* **2024**.
129. Mhatre, A. Detecting the presence of social bias in GPT-3.5 using association tests. In Proceedings of the 2023 international conference on advanced computing technologies and applications (ICACTA). IEEE, 2023, pp. 1–6.
130. Yang, Y.; Duan, H.; Abbasi, A.; Lalor, J.P.; Tam, K.Y. Bias a-head? analyzing bias in transformer-based language model attention heads. In Proceedings of the Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025), 2025, pp. 276–290.
131. Ma, S.; Salinas, A.; Nyarko, J.; Henderson, P. Breaking Down Bias: On The Limits of Generalizable Pruning Strategies. In Proceedings of the Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, 2025, pp. 2437–2450.
132. Zayed, A.; Mordido, G.; Shabanian, S.; Baldini, I.; Chandar, S. Fairness-aware structured pruning in transformers. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 22484–22492.
133. Hossain, P.S.; Raj, C.; Zhu, Z.; Lin, J.; Marasco, E. Toward Inclusive Language Models: Sparsity-Driven Calibration for Systematic and Interpretable Mitigation of Social Biases in LLMs. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025, 2025, pp. 2475–2508.
134. Lundberg, S. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* **2017**.
135. Ribeiro, M.T.; Singh, S.; Guestrin, C. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
136. Devine, P.G. Stereotypes and Prejudice: Their Automatic and Controlled Components. *J. Personal. Soc. Psychol.* **1989**, *56*, 5–18. https://doi.org/10.1037/0022-3514.56.1.5.
137. Snyder, M.; Swann, W.B. Hypothesis-testing processes in social interaction. *J. Personal. Soc. Psychol.* **1978**, *36*, 1202–1212. https://doi.org/10.1037/0022-3514.36.11.1202.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0815.v1

138. Dovidio, J.F.; Hewstone, M.; Glick, P.; Esses, V.M. Prejudice, Stereotyping and Discrimination: Theoretical and Empirical Overview. In *The SAGE Handbook of Prejudice, Stereotyping and Discrimination*; Dovidio, J.F.; Hewstone, M.; Glick, P.; Esses, V.M., Eds.; SAGE Publications, 2010; pp. 3–28.
139. Bodenhausen, G.V.; Richeson, J.A. Prejudice, Stereotyping, and Discrimination. In *Advanced Social Psychology: The State of the Science*; Baumeister, R.F.; Finkel, E.J., Eds.; Oxford University Press, 2010; pp. 350–380.
140. Allport, G.W. *The Nature of Prejudice*; Addison-Wesley: Reading, MA, 1954.
141. Pettigrew, T.F. Intergroup Contact Theory. *Annu. Rev. Psychol.* **1998**, *49*, 65–85. https://doi.org/10.1146/annurev.psych.49.1.65.
142. Becker, G.S. *The Economics of Discrimination*; University of Chicago Press: Chicago, 1957.
143. Phelps, E.S. The Statistical Theory of Racism and Sexism. *Am. Econ. Rev.* **1972**, *62*, 659–661.
144. Arrow, K.J. The Theory of Discrimination. In *Discrimination in Labor Markets*; Ashenfelter, O.; Rees, A., Eds.; Princeton University Press: Princeton, NJ, 1973; pp. 3–33.
145. Pager, D.; Shepherd, H. The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets. *Annu. Rev. Sociol.* **2008**, *34*, 181–209.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.