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

Anonymous ACL submission

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.

1 Introduction

Humans are quite good at identifying patterns and forming clusters. They naturally construct conceptual groupings based on shared features, even under uncertainty (Bruner et al., 1956; Shepard et al., 1961), though this process does not always follow strictly logical rules (Rosch, 2024). Humans tend to classify those similar to themselves as the "in-group" and those perceived as different as the "out-group" (Brewer, 1999; Linville et al., 1989; Mullen et al., 1992; Fiske et al., 2002). Theories such as the similarity–attraction hypothesis (Byrne, 1971) and social identity theory (Tajfel and Turner, 1979) 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 (Brewer, 1999; Fiske et al., 2002; Turner and Reynolds, 2003; Cuddy et al., 2004, 2008). 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] (Harari, 2014; Mobbs et al., 2015; LeDoux, 2012; Wise, 2009). Social competition and interpersonal relationships likewise constitute fundamental survival-relevant contexts (Lakoff, 2024; Rosenbaum, 2014; Griffin, 2001), 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 (Liu et al., 2025) to medical settings, including clinical report generation (Busch et al., 2025), 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 (Pagano et al., 2023; Jeoung et al., 2023; Guo et al., 2024). To mitigate these challenges, research has initially focused on assessing bias in LLMs by exploiting var-

[1] System 1 refers to fast, automatic, intuitive, and emotion-driven cognition, in contrast to System 2, which is slower, deliberate, and analytical (Kahneman, 2011).

ious tasks, including Natural Language Generation (NLG) (Nadeem et al., 2021; Felkner et al., 2023), counterfactual reasoning (Nangia et al., 2020; Sahoo et al., 2024), Natural Language Inference (NLI) (Baldini et al., 2023), Question Answering (Parrish et al., 2022; Tomar et al., 2025b), and prompt completion (Gehman et al., 2020; Dhamala et al., 2021). Gallegos et al. (2024) 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 (Shejole and Bhattacharyya, 2025). 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. (2025a) 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 (§ 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 (§ 3).

2. A multimodal, linguistic, and geographic analysis of stereotype research, identifying key gaps and underexplored requirements (§ 4).

3. A unified analysis of challenges in stereotype research by integrating social-psychological and computational perspectives (§ 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 (§ 6).

2 Social Psychological Perspectives on Stereotypes

In this section, we review key social psychological theories (§ 2.1) and frameworks (§ 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 (Byrne, 1971) and Social Identity Theory (Tajfel and Turner, 1979) posit systematic *in-group* favoritism, whereby individuals favor in-groups over out-groups to enhance self-esteem (Turner and Reynolds, 2003). 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 (Ellemers and Haslam, 2012; Postmes and Branscombe, 2010). 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 (Eagly, 1987) 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 aware of its influence (Ward and Friedman, 2006). 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 (Gauntlett, 2008; Bartlett et al., 2013). Social Role Theory is closely related to Social Learning Theory (Bandura and Walters, 1977), 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 funda-

mental as individual-based perception (Turner et al., 1987). 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*** (Augoustinos and Walker, 1998), grounded in group-based cognition.

4. *Theories Discussing Social Cognition:* Social cognition–based theories (Fiske, 1992; Fiske et al., 1993; Fiske and Haslam, 1996; Fiske and Taylor, 2020) 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 (Jost et al., 2004; Jost and Van der Toorn, 2012; Jost, 2019) 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 (Banaji, 2002). 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 (1992) 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 (1991) conceptualize stereotypes and categorization as ***discursive constructions*** rather than cognitive processes (Augoustinos and Walker, 1998).

7. *Intersectionality Theory:* Recent work (Cho et al., 2013; Carastathis, 2014; Crenshaw, 2013) 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) (Fiske et al., 2002). Warmth judgments are shaped primarily by perceived competition, while competence judgments reflect perceived status. These dimensions yield four canonical stereotype profiles: 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 (Fiske et al., 2002; Cuddy et al., 2011).

2. *Agency–Beliefs–Communion (ABC) Model:* The ABC model[2] (Koch et al., 2016) 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. (2002) considers competence as Agency (A) and warmth as Communion (C). Abele et al. (2016) 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. (2025) and Shejole and Bhattacharyya (2025) 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 (Turner et al., 1987). 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 3 (Appendix C) summarizes these theories and frameworks and Table 2 (Appendix C) 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. (2021) computationally operationalized the SCM by deriving warmth and competence directions from lexicon-based word embeddings (Nicolas et al., 2021) 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. (2022) 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 (Herold et al., 2022) and bias mitigation (Ungless et al., 2022; Omrani et al., 2023). Cao et al. (2022) operationalized the ABC model as a computational framework to identify group–trait associations in language models, demonstrating moderate alignment with human judgments, supporting intersectional analysis, and evaluating the

[2]The terms Agency (A) and Communion (C) were coined by Bakan (1966).

[3]Anti-stereotypes refer to attributes strongly counter to commonly held beliefs about a social group (e.g., football players being weak).

approach in a U.S.-centric context. These works underscores the multidimensional structure of stereotypes. Building on this view, Fraser et al. (2024) 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 (2025) 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 (Eagly, 1987) 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 (Schweinitz et al., 2010; Kumar et al., 2022; Xu et al., 2019; Gallego et al., 2025; Shehatta, 2020; Atillah et al., 2020; Ji, 2021; Madaan et al., 2017a,b, 2018). More recently, Wang and Lin (2024) 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 (Lelwica, 2011), whereas medium-sized bodies are more socially preferred in some Middle Eastern contexts (Khalaf et al., 2015; Musaiger et al., 2000). 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 (Nadeem et al., 2021), CrowS-Pairs (Nangia et al., 2020), and BBQ (Parrish et al., 2021) 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 (Asad et al., 2025) 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 (2024) reveal substantial gender and racial biases in large vision-language models (VLMs), while Jha et al. (2024) 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 (Lee et al., 2025; Zhou et al., 2022; Pang, 2025; Srinivasan and Bisk, 2022; Hamidieh et al., 2024; Malik and Johansson, 2022) further confirms the prevalence of stereotypical biases in VLMs, underscoring a critical and underexplored challenge for multimodal AI. In contrast, research on speech and video remains limited; for example, Kurinec and Weaver III (2021) show that vocal cues alone can activate racial stereotypes. These findings highlight the need for broader investigation into stereotype detection and mitigation in

[4]These dimensions were Sociability, Morality, Ability, Assertiveness, Beliefs, and Status.

[5]Skin complexion, body shape, height, attire, hair texture, and eye color.

conversational audio and video modalities.

4.2 Linguistic and Geographic Coverage

SeeGULL (Jha et al., 2023) and *Visage* (Jha et al., 2024) examine geographic variation in stereotypes, but primarily operationalize geography through nationality. *WinoQueer* (Felkner et al., 2023) focuses on stereotypes related to LGBTQ+ identities, providing a dedicated resource for studying sexual and gender minority representation. Benchmarks such as *EMGSD* (King et al., 2024) and *MGSD* (Zekun et al., 2023), inspired by earlier bias datasets including *StereoSet* (Nadeem et al., 2021) and *CrowS-Pairs* (Nangia et al., 2020), 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 (Blodgett et al., 2021). *StereoDetect* (Shejole and Bhattacharyya, 2025) 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* (Jin et al., 2024), *KOLD* (Jeong et al., 2022)), *French* (*French-CrowS-Pairs* (Névéol et al., 2022)), *Hindi* (*IndiBias* (Sahoo et al., 2024), *BharatBBQ* (Tomar et al., 2025b)), and *Italian* (*FB-Stereotypes* Bosco et al., 2023, *QueeroTypes* (Cignarella et al., 2024), *StereoHoax-IT* (Schmeisser-Nieto et al., 2024)). The multilingual *MRHC* dataset (Bourgeade et al., 2023), covering Italian, Spanish, and French, examines racial stereotypes in social media. More recently, *SHADES* (Mitchell et al., 2025) 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 (2025) 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 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.

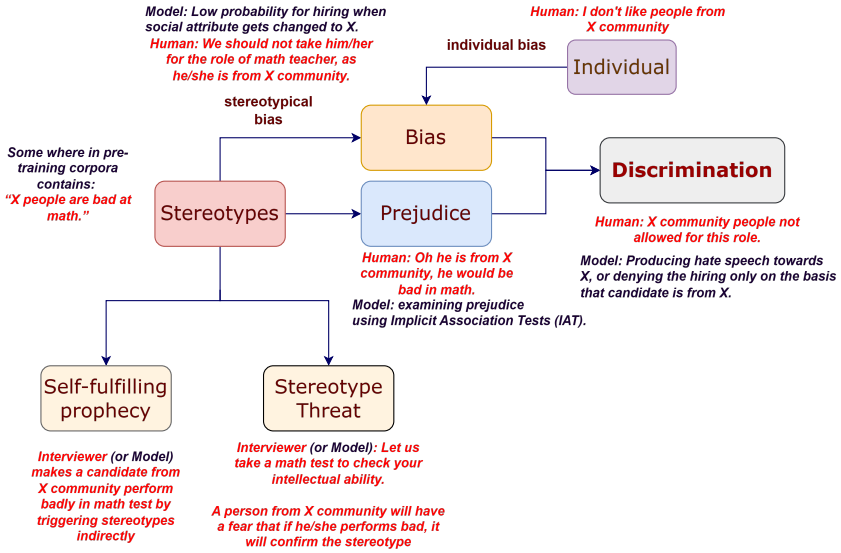

Figure 1: Inter-relationship between of concepts of social psychology and connecting it with Responsible AI scenarios.

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. (2025); Nie et al. (2024); Gamboa et al. (2025). 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. (2025) 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 (1995) 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 (Shrawgi et al., 2024). 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*** (Merton, 1948; Jussim, 1986). 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). Discrim-

ination is the most evident, with numerous computational studies demonstrating biased behavior in hiring scenarios (Peña et al., 2025; Anzenberg et al., 2025; Wang et al., 2024; An et al., 2024; Armstrong et al., 2024). Further research is needed to determine whether LLMs and AI models exhibit personal biases similar to humans and to understand the underlying causes.

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 (Greenwald and Krieger, 2006), is a key contributor to prejudice (Kahn, 2017; Payne et al., 2017) and is shaped by automatic cognitive processes, as described in Social Cognition Theory (Section 2). The Implicit Association Test (IAT) (Greenwald et al., 1998, 2009) was developed to measure this phenomenon. Similar tests applied to LLMs (Bai et al., 2024; Mhatre, 2023) 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 D. 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 (Cuddy et al., 2008; Fraser et al., 2021) 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 (Yang et al., 2025; Ma et al., 2025; Zayed et al., 2024; Hossain et al., 2025), 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 (Lundberg, 2017) and LIME (Ribeiro et al., 2016) 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 Conclusion

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, 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 (Appendix B). 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.

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.

References

Andrea E Abele, Nicole Hauke, Kim Peters, Eva Louvet, Aleksandra Szymkow, and Yanping Duan. 2016. Facets of the fundamental content dimensions: Agency with competence and assertiveness—communion with warmth and morality. *Frontiers in psychology*, 7:1810.

Gordon W. Allport. 1954. *The Nature of Prejudice*. Addison-Wesley, Reading, MA.

Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. 2024. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? *arXiv preprint arXiv:2406.10486*.

Eitan Anzenberg, Arunava Samajpati, Sivasankaran Chandrasekar, and Varun Kacholia. 2025. Evaluating the promise and pitfalls of llms in hiring decisions. *arXiv preprint arXiv:2507.02087*.

Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. 2024. The silicon ceiling: Auditing gpt's race and gender biases in hiring. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–18.

Kenneth J. Arrow. 1973. The theory of discrimination. In Orley Ashenfelter and Albert Rees, editors, *Discrimination in Labor Markets*, pages 3–33. Princeton University Press, Princeton, NJ.

Narjis Asad, Nihar Ranjan Sahoo, Rudra Murthy, Swaprava Nath, and Pushpak Bhattacharyya. 2025. "you are beautiful, body image stereotypes are ugly!" BIStereo: A benchmark to measure body image stereotypes in language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24471–24496, Vienna, Austria. Association for Computational Linguistics.

Widya Atillah, M Bahri Arifin, and Nita Maya Valiantien. 2020. An analysis of stereotype in zootopia movie. *Ilmu Budaya: Jurnal Bahasa, Sastra, Seni, dan Budaya*, 4(1):49–62.

Martha Augoustinos and Iain Walker. 1998. The construction of stereotypes within social psychology: From social cognition to ideology. *Theory & Psychology*, 8(5):629–652.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. 2024. Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*.

David Bakan. 1966. The duality of human existence: An essay on psychology and religion.

Ioana Baldini, Chhavi Yadav, Manish Nagireddy, Payel Das, and Kush R Varshney. 2023. Keeping up with the language models: Systematic benchmark extension for bias auditing. *arXiv preprint arXiv:2305.12620*.

Mahzarin R Banaji. 2002. Stereotypes, social psychology of. *International encyclopedia of the social and behavioral sciences*, pages 15100–15104.

Albert Bandura and Richard H Walters. 1977. *Social learning theory*, volume 1. Prentice hall Englewood Cliffs, NJ.

Djurdja Bartlett, Agnes Rocamora, and Shaun Cole. 2013. Fashion media.

Gary S. Becker. 1957. *The Economics of Discrimination*. University of Chicago Press, Chicago.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *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)*, pages 1004–1015, Online. Association for Computational Linguistics.

Galen V. Bodenhausen and Jennifer A. Richeson. 2010. Prejudice, stereotyping, and discrimination. In Roy F. Baumeister and Eli J. Finkel, editors, *Advanced Social Psychology: The State of the Science*, pages 350–380. Oxford University Press.

Cristina Bosco, Viviana Patti, Simona Frenda, Alessandra Teresa Cignarella, Marinella Paciello, and Francesca D'Errico. 2023. Detecting racial stereotypes: An italian social media corpus where psychology meets nlp. *Information Processing & Management*, 60(1):103118.

Tom Bourgeade, Alessandra Teresa Cignarella, Simona Frenda, Mario Laurent, Wolfgang Schmeisser-Nieto, Farah Benamara, Cristina Bosco, Véronique Moriceau, Viviana Patti, and Mariona Taulé. 2023. A multilingual dataset of racial stereotypes in social media conversational threads. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 686–696.

Marilynn B Brewer. 1999. The psychology of prejudice: Ingroup love and outgroup hate? *Journal of social issues*, 55(3):429–444.

Jerome S Bruner, Jacqueline J Goodnow, and A George. 1956. Austin. a study of thinking. *New York: John Wiley & Sons, Inc*, 14:330.

Felix Busch, Lena Hoffmann, Daniel Pinto Dos Santos, Marcus R Makowski, Luca Saba, Philipp Prucker, Martin Hadamitzky, Nassir Navab, Jakob Nikolas Kather, Daniel Truhn, and 1 others. 2025. Large language models for structured reporting in radiology: past, present, and future. *European Radiology*, 35(5):2589–2602.

D Byrne. 1971. The attraction paradigm academic press. *New York, NY, USA*.

Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of us social stereotypes in english language models. *arXiv preprint arXiv:2206.11684*.

Anna Carastathis. 2014. The concept of intersectionality in feminist theory. *Philosophy compass*, 9(5):304–314.

Sumi Cho, Kimberlé Williams Crenshaw, and Leslie McCall. 2013. Toward a field of intersectionality studies: Theory, applications, and praxis. *Signs: Journal of women in culture and society*, 38(4):785–810.

Alessandra Teresa Cignarella, Manuela Sanguinetti, Simona Frenda, Andrea Marra, Cristina Bosco, and Valerio Basile. 2024. Queereotypes: A multi-source italian corpus of stereotypes towards lgbtqia+ community members. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13429–13441.

Kimberlé Crenshaw. 2013. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*, pages 23–51. Routledge.

Amy J. C. Cuddy, Susan T. Fiske, and Peter Glick. 2011. Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, 50(3):472–486.

Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2004. When professionals become mothers, warmth doesn't cut the ice. *Journal of Social issues*, 60(4):701–718.

Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in experimental social psychology*, 40:61–149.

Aida Mostafazadeh Davani, Sunipa Dev, Héctor Pérez-Urbina, and Vinodkumar Prabhakaran. 2025. A comprehensive framework to operationalize social stereotypes for responsible ai evaluations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30018–30031.

Patricia G. Devine. 1989. Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, 56(1):5–18.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

John F. Dovidio, Miles Hewstone, Peter Glick, and Victoria M. Esses. 2010. Prejudice, stereotyping and discrimination: Theoretical and empirical overview. In John F. Dovidio, Miles Hewstone, Peter Glick, and Victoria M. Esses, editors, *The SAGE Handbook of Prejudice, Stereotyping and Discrimination*, pages 3–28. SAGE Publications.

Alice H. Eagly. 1987. *Sex Differences in Social Behavior: A Social-role Interpretation*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Derek Edwards. 1991. Categories are for talking: On the cognitive and discursive bases of categorization. *Theory & psychology*, 1(4):515–542.

Naomi Ellemers and S Alexander Haslam. 2012. Social identity theory. *Handbook of theories of social psychology*, 2:379–398.

Virginia K Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. *arXiv preprint arXiv:2306.15087*.

Alan P Fiske and Nick Haslam. 1996. Social cognition is thinking about relationships. *Current directions in psychological science*, 5(5):143–148.

Susan T Fiske. 1992. Thinking is for doing: portraits of social cognition from daguerreotype to laserphoto. *Journal of personality and social psychology*, 63(6):877.

Susan T. Fiske, Amy J. C. Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6):878–902.

Susan T Fiske and 1 others. 1993. Social cognition and social perception. *Annual review of psychology*, 44(1):155–194.

Susan T Tufts Fiske and Shelley E Taylor. 2020. Social cognition: From brains to culture.

Kathleen Fraser and Svetlana Kiritchenko. 2024. Examining gender and racial bias in large vision–language models using a novel dataset of parallel images. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 690–713, St. Julian's, Malta. Association for Computational Linguistics.

Kathleen Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. 2024. How does stereotype content differ across data sources? In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 18–34, Mexico City, Mexico. Association for Computational Linguistics.

Kathleen C Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. 2022. Computational modeling of stereotype content in text. *Frontiers in artificial intelligence*, 5:826207.

Kathleen C Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. *arXiv preprint arXiv:2106.02596*.

Ana Guadalupe Gallego, Camino Ferreira, and Ana Rosa Arias-Gago. 2025. Stereotyped representations of disability in film and television: A critical review of narrative media. *Disabilities*, 5(4):1–25.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Lance Calvin Lim Gamboa, Yue Feng, and Mark Lee. 2025. Social bias in multilingual language models: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27845–27868.

David Gauntlett. 2008. *Media, gender and identity: An introduction*. Routledge.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369.

Anthony G Greenwald and Linda Hamilton Krieger. 2006. Implicit bias: Scientific foundations. *California law review*, 94(4):945–967.

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464–1480.

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 2009. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464–1480.

Donald R Griffin. 2001. *Animal minds: Beyond cognition to consciousness*. University of Chicago Press.

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915*.

Kimia Hamidieh, Haoran Zhang, Walter Gerych, Thomas Hartvigsen, and Marzyeh Ghassemi. 2024. Identifying implicit social biases in vision-language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 547–561.

Yuval Noah Harari. 2014. *Sapiens: A brief history of humankind*. Random House.

Brienna Herold, James Waller, and Raja Kushalnagar. 2022. Applying the stereotype content model to assess disability bias in popular pre-trained nlp models underlying ai-based assistive technologies. In *Ninth workshop on speech and language processing for assistive technologies (SLPAT-2022)*, pages 58–65.

Prommy Sultana Hossain, Chahat Raj, Ziwei Zhu, Jessica Lin, and Emanuela Marasco. 2025. Toward inclusive language models: Sparsity-driven calibration for systematic and interpretable mitigation of social biases in llms. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2475–2508.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. Kold: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833.

Sullam Jeoung, Yubin Ge, and Jana Diesner. 2023. StereoMap: Quantifying the awareness of human-like stereotypes in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12236–12256, Singapore. Association for Computational Linguistics.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan Reddy, and Sunipa Dev. 2024. Visage: A global-scale analysis of visual stereotypes in text-to-image generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12333–12347.

Jiaxin Ji. 2021. Analysis of gender stereotypes in disney female characters. In *2021 3rd International Conference on Literature, Art and Human Development (ICLAHD 2021)*, pages 451–454. Atlantis Press.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. Kobbq: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

John T Jost. 2019. A quarter century of system justification theory: Questions, answers, criticisms, and societal applications. *British Journal of Social Psychology*, 58(2):263–314.

John T Jost, Mahzarin R Banaji, and Brian A Nosek. 2004. A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political psychology*, 25(6):881–919.

John T Jost and Jojanneke Van der Toorn. 2012. System justification theory. *Handbook of theories of social psychology*, 2:313–343.

Lee Jussim. 1986. Self-fulfilling prophecies: A theoretical and integrative review. *Psychological review*, 93(4):429.

Jonathan Kahn. 2017. Pills for prejudice: implicit bias and technical fix for racism. *American Journal of Law & Medicine*, 43(2-3):263–278.

Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux*.

Atika Khalaf, Albert Westergren, Vanja Berggren, Örjan Ekblom, and Hazzaa M. Al-Hazzaa. 2015. Perceived and ideal body image in young women in south western saudi arabia. *Journal of Obesity*, 2015(1):697163.

Michelle YoungJin Kim and Kristen Johnson. 2025. Korean stereotype content model: Translating stereotypes across cultures. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 59–70.

Theo King, Zekun Wu, Adriano Koshiyama, Emre Kazim, and Philip Treleaven. 2024. Hearts: A holistic framework for explainable, sustainable and robust text stereotype detection. *arXiv preprint arXiv:2409.11579*.

Alex Koch, Roland Imhoff, Ron Dotsch, Christian Unkelbach, and Hans Alves. 2016. The abc of stereotypes about groups: Agency/socioeconomic success, conservative-progressive beliefs, and communion. *Journal of Personality and Social Psychology*, 110(5):675–709.

Arjun M Kumar, Jasmine YQ Goh, Tiffany HH Tan, and Cynthia SQ Siew. 2022. Gender stereotypes in hollywood movies and their evolution over time: Insights from network analysis. *Big Data and Cognitive Computing*, 6(2):50.

Courtney A Kurinec and Charles A Weaver III. 2021. "sounding black": Speech stereotypicality activates racial stereotypes and expectations about appearance. *Frontiers in psychology*, 12:785283.

George Lakoff. 2024. *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press.

Joseph LeDoux. 2012. Rethinking the emotional brain. *Neuron*, 73(4):653–676.

Messi HJ Lee, Soyeon Jeon, Jacob M Montgomery, and Calvin K Lai. 2025. Visual cues of gender and race are associated with stereotyping in vision-language models. *arXiv preprint arXiv:2503.05093*.

Michelle Lelwica. 2011. The religion of thinness. *Scripta Instituti Donneriani Aboensis*, 23:257–285.

Patricia W Linville, Gregory W Fischer, and Peter Salovey. 1989. Perceived distributions of the characteristics of in-group and out-group members: empirical evidence and a computer simulation. *Journal of personality and social psychology*, 57(2):165.

Jiayi Liu, Bo Jiang, and Yu'ang Wei. 2025. Llms as promising personalized teaching assistants: How do they ease teaching work? *ECNU Review of Education*, 8(2):343–348.

Scott Lundberg. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Sibo Ma, Alejandro Salinas, Julian Nyarko, and Peter Henderson. 2025. Breaking down bias: On the limits of generalizable pruning strategies. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 2437–2450.

Nishtha Madaan, Sameep Mehta, Taneea Agrawaal, Vrinda Malhotra, Aditi Aggarwal, Yatin Gupta, and Mayank Saxena. 2018. Analyze, detect and remove gender stereotyping from bollywood movies. In *Conference on fairness, accountability and transparency*, pages 92–105. PMLR.

Nishtha Madaan, Sameep Mehta, Taneea S Agrawaal, Vrinda Malhotra, Aditi Aggarwal, and Mayank Saxena. 2017a. Analyzing gender stereotyping in bollywood movies. *arXiv preprint arXiv:1710.04117*.

Nishtha Madaan, Sameep Mehta, Mayank Saxena, Aditi Aggarwal, Taneea S Agrawaal, and Vrinda Malhotra. 2017b. Bollywood movie corpus for text, images and videos. *arXiv preprint arXiv:1710.04142*.

Manuj Malik and Richard Johansson. 2022. Controlling for stereotypes in multimodal language model evaluation. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 263–271, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Robert K Merton. 1948. The self-fulfilling prophecy. *The antioch review*, 8(2):193–210.

Aatmaj Mhatre. 2023. Detecting the presence of social bias in gpt-3.5 using association tests. In *2023 international conference on advanced computing technologies and applications (ICACTA)*, pages 1–6. IEEE.

Margaret Mitchell, John Smith, Alice Lee, Ravi Kumar, and Li Wang. 2025. Shades: Towards a multilingual assessment of stereotypes in language models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11995–12041. Association for Computational Linguistics.

Dean Mobbs, Cindy C Hagan, Tim Dalgleish, Brian Silston, and Charlotte Prévost. 2015. The ecology of human fear: survival optimization and the nervous system. *Frontiers in neuroscience*, 9:121062.

Brian Mullen, John F Dovidio, Craig Johnson, and Carolyn Copper. 1992. In-group-out-group differences in social projection. *Journal of Experimental Social Psychology*, 28(5):422–440.

Abdulrahman O Musaiger, Abdul-hai A Al-Awadi, and Mariam A Al-Mannai. 2000. Lifestyle and social factors associated with obesity among the bahraini adult population. *Ecology of food and nutrition*, 39(2):121–133.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *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)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. 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 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531.

Gandalf Nicolas, Xuechunzi Bai, and Susan T Fiske. 2021. Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, 51(1):178–196.

Shangrui Nie, Michael Fromm, Charles Welch, Rebekka Görge, Akbar Karimi, Joan Plepi, Nazia Mowmita, Nicolas Flores-Herr, Mehdi Ali, and Lucie Flek. 2024. Do multilingual large language models mitigate stereotype bias? In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 65–83.

Ali Omrani, Alireza Salkhordeh Ziabari, Charles Yu, Preni Golazizian, Brendan Kennedy, Mohammad Atari, Heng Ji, and Morteza Dehghani. 2023. Social-group-agnostic bias mitigation via the stereotype content model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4123–4139.

Tiago P Pagano, Rafael B Loureiro, Fernanda VN Lisboa, Rodrigo M Peixoto, Guilherme AS Guimarães, Gustavo OR Cruz, Maira M Araujo, Lucas L Santos, Marco AS Cruz, Ewerton LS Oliveira, and 1 others. 2023. Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, 7(1):15.

Devah Pager and Hana Shepherd. 2008. The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annual Review of Sociology*, 34:181–209.

Bo Pang. 2025. *Investigating Stereotypical Bias in Large Language and Vision-Language Models*. Ph.D. thesis, University of Auckland New Zealand.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2021. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

B Keith Payne, Heidi A Vuletich, and Kristjen B Lundberg. 2017. The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, 28(4):233–248.

Alejandro Peña, Julian Fierrez, Aythami Morales, Gonzalo Mancera, Miguel Lopez-Duran, and Ruben Tolosana. 2025. Addressing bias in llms: Strategies and application to fair ai-based recruitment. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 1976–1987.

Thomas F. Pettigrew. 1998. Intergroup contact theory. *Annual Review of Psychology*, 49:65–85.

Thomas F. Pettigrew and Linda R. Tropp. 2006. How does intergroup contact reduce prejudice? meta-analytic tests of three mediators.

Edmund S. Phelps. 1972. The statistical theory of racism and sexism. *The American Economic Review*, 62(4):659–661.

Tom Ed Postmes and Nyla R Branscombe. 2010. *Rediscovering social identity.* Psychology Press.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Eleanor Rosch. 2024. Principles of categorization. In *Cognition and categorization*, pages 27–48. Routledge.

David A Rosenbaum. 2014. *It's a jungle in there: How competition and cooperation in the brain shape the mind.* Oxford University Press.

Nihar Sahoo, Pranamya Kulkarni, Arif Ahmad, Tanu Goyal, Narjis Asad, Aparna Garimella, and Pushpak Bhattacharyya. 2024. IndiBias: A benchmark dataset to measure social biases in language models for Indian context. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8786–8806, Mexico City, Mexico. Association for Computational Linguistics.

Wolfgang S Schmeisser-Nieto, Alessandra Teresa Cignarella, Tom Bourgeade, Simona Frenda, Alejandro Ariza-Casabona, Laurent Mario, Paolo Giovanni Cicirelli, Andrea Marra, Giuseppe Corbelli, Farah Benamara, and 1 others. 2024. Stereohoax: a multilingual corpus of racial hoaxes and social media reactions annotated for stereotypes. *Language Resources and Evaluation*, pages 1–39.

Jörg Schweinitz, Johanna Eder, Fotis Jannidis, and Ralf Schneider. 2010. Stereotypes and the narratological analysis of film characters. *Revisionen*, (3):276–289.

Sarah Shehatta. 2020. Breaking stereotypes: A multimodal analysis of the representation of the female lead in the animation movie brave. *Textual Turnings: An International Peer-Reviewed Journal in English Studies*, 2(1):170–194.

Kaustubh Shivshankar Shejole and Pushpak Bhattacharyya. 2025. StereoDetect: Detecting stereotypes and anti-stereotypes the correct way using social psychological underpinnings. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4051–4082, Suzhou, China. Association for Computational Linguistics.

Roger N Shepard, Carl I Hovland, and Herbert M Jenkins. 1961. Learning and memorization of classifications. *Psychological monographs: General and applied*, 75(13):1.

Hari Shrawgi, Prasanjit Rath, Tushar Singhal, and Sandipan Dandapat. 2024. Uncovering stereotypes in large language models: A task complexity-based approach. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1841–1857. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799.

Mark Snyder and William B. Swann. 1978. Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology*, 36(11):1202–1212.

Tejas Srinivasan and Yonatan Bisk. 2022. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 77–85.

Claude M. Steele and Joshua Aronson. 1995. Stereotype threat and the intellectual test performance of african americans. *Journal of Personality and Social Psychology*, 69(5):797–811.

Henri Tajfel and John C. Turner. 1979. An integrative theory of intergroup conflict. In William G. Austin and Stephen Worchel, editors, *The Social Psychology of Intergroup Relations*, pages 33–47. Brooks/Cole, Monterey, CA.

Aditya Tomar, Rudra Murthy, and Pushpak Bhattacharyya. 2025a. Stereotype detection as a catalyst for enhanced bias detection: A multi-task learning approach. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17304–17317, Vienna, Austria. Association for Computational Linguistics.

Aditya Tomar, Nihar Ranjan Sahoo, and Pushpak Bhattacharyya. 2025b. Bharatbbq: A multilingual bias benchmark for question answering in the indian context. *Transactions of the Association for Computational Linguistics*, 13:1672–1692.

John C Turner, Michael A Hogg, Penelope J Oakes, Stephen D Reicher, and Margaret S Wetherell. 1987. *Rediscovering the social group: A self-categorization theory*. basil Blackwell.

John C Turner and Katherine J Reynolds. 2003. The social identity perspective in intergroup relations: Theories, themes, and controversies. *Blackwell handbook of social psychology: Intergroup processes*, pages 133–152.

Eddie Ungless, Amy Rafferty, Hrichika Nag, and Björn Ross. 2022. A robust bias mitigation procedure based on the stereotype content model. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS)*, pages 207–217.

Yilin Wang and Chujun Lin. 2024. Stereotypes at the intersection of perceivers, situations, and identities: analyzing stereotypes from storytelling using natural language processing.

Ze Wang, Zekun Wu, Xin Guan, Michael Thaler, Adriano Koshiyama, Skylar Lu, Sachin Beepath, Ediz Ertekin, and Maria Perez-Ortiz. 2024. Jobfair: A framework for benchmarking gender hiring bias in large language models. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 3227–3246.

L Monique Ward and Kimberly Friedman. 2006. Using tv as a guide: Associations between television viewing and adolescents' sexual attitudes and behavior. *Journal of research on adolescence*, 16(1):133–156.

Margaret Wetherell and Jonathan Potter. 1992. *Mapping the language of racism: Discourse and the legitimation of exploitation*. Columbia University Press.

Jeff Wise. 2009. *Extreme fear: The science of your mind in danger*. St. Martin's Press.

Huimin Xu, Zhang Zhang, Lingfei Wu, and Cheng-Jun Wang. 2019. The cinderella complex: Word embeddings reveal gender stereotypes in movies and books. *PloS one*, 14(11):e0225385.

Yi Yang, Hanyu Duan, Ahmed Abbasi, John P Lalor, and Kar Yan Tam. 2025. Bias a-head? analyzing bias in transformer-based language model attention heads. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 276–290.

Abdelrahman Zayed, Gonçalo Mordido, Samira Shabanian, Ioana Baldini, and Sarath Chandar. 2024. Fairness-aware structured pruning in transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22484–22492.

Wu Zekun, Sahan Bulathwela, and Adriano Soares Koshiyama. 2023. Towards auditing large language models: Improving text-based stereotype detection. *ArXiv*, abs/2311.14126.

Kankan Zhou, Eason Lai, and Jing Jiang. 2022. Vlstereoset: A study of stereotypical bias in pre-trained vision-language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 527–538.

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.

A.1 Stereotypes

Stereotypes are overgeneralized beliefs about members of a social category, attributing uniform traits to all individuals and ignoring individual differences (Devine, 1989). For example, "Asians are good at math" overlooks variation within the group and can lead to unfair assumptions (Snyder and Swann, 1978; Steele and Aronson, 1995). Anti-stereotypes—beliefs that counter prevailing stereotypes—can reduce biased thinking (Devine, 1989). Only beliefs about social categories, not general truths, qualify as stereotypes.

A.2 Bias

Bias refers to inclinations or partiality favoring or disadvantaging certain groups, which can be explicit (conscious) or implicit (automatic) (Dovidio et al., 2010; Bodenhausen and Richeson, 2010). Explicit bias underlies overt discrimination, whereas implicit bias operates unconsciously, subtly influencing perceptions and behaviors (Fiske et al., 2002). Bias is distinct from stereotypes, though stereotypical biases arise from underlying stereotypical beliefs (Gallegos et al., 2024).

A.3 Prejudice

Prejudice is an affective attitude toward individuals based solely on their social category, reflecting emotions such as fear, contempt, or dislike (Allport, 1954; Devine, 1989). It often arises automatically (System 1) but can be mitigated through deliberate reflection (System 2) (Kahneman, 2011). Prejudice forms the emotional basis for discriminatory behavior and can be reduced by positive intergroup contact (Pettigrew and Tropp, 2006; Pettigrew, 1998).

A.4 Discrimination

Discrimination is the behavioral enactment of biased attitudes, leading to unfair treatment of individuals or groups (Allport, 1954). It can be:

- **Direct**: overt actions such as refusing service or workplace harassment (Becker, 1957; Dovidio et al., 2010).
- **Indirect**: neutral-appearing policies or practices that disproportionately disadvantage certain groups, e.g., standardized tests or institutional barriers (Phelps, 1972; Arrow, 1973; Pager and Shepherd, 2008).

A.5 Distinguishing Stereotypes from Bias

Recent work (Shejole and Bhattacharyya, 2025) highlights persistent conceptual confusion between stereotypes and biases, which has led to the construction of benchmarks having inconsistenices for stereotypes (e.g., MGSD (Zekun et al., 2023), EMGSD King et al. (2024)). This confusion limits the validity and generalizability of stereotype-detection models, as they often capture surface-level biases rather than the underlying social structures 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. (2024).

B Summarizing Future Directions

In Section 7, we argued that grounding future computational research in established social psychological foundations, together with the research directions outlined in this paper, can enable the development of more principled, culturally grounded, and effective Responsible AI interventions. We summarize these future research directions in Table 1. The table provides a systematic synthesis of the research scope and open opportunities identified throughout this review, explicitly mapping them to the paper's sections and subsections. It highlights key directions for bridging social psychological theories, such as the Five-Tuple Framework, with computational research areas including multimodal narrative analysis and broader global linguistic coverage. In addition, the table identifies challenges related to scalability and the evolving nature of stereotypes, while situating these issues within Responsible AI efforts focused on implicit bias detection and model interpretability. By organizing these gaps and opportunities, Table 1 offers a structured roadmap for future interdisciplinary work aimed at understanding and mitigating stereotypes in LLMs.

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

Section	Subsection	Future Research Scope & Opportunities
Section 2	Major Frameworks (§ 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 (§ 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 (§ 3.2)	Implement proactive identification of stereotypes in media narratives to assess and mitigate potential social harms before dissemination.
	Body-Image (§ 3.3)	Systematically quantify body-image bias in LLMs and develop automatic modeling from media representations to monitor stereotypical ideals.
Section 4	Multimodality (§ 4.1)	Expand investigations into stereotype detection and mitigation beyond text and images to include conversational audio and video.
	Linguistic/Geographic Coverage (§ 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 (§ 5.1)	Research more efficient methods for social analysis to help models handle unseen target groups and extract context-specific information.
	Annotation (§ 5.2)	Select representative annotator subsets reflecting the target community to ensure unbiased benchmarks and avoid skewed selections.
	Scalability (§ 5.3)	Explore strategies for modeling contexts separately to achieve global inclusivity despite current resource and scalability constraints.
	Dynamic Nature (§ 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 (§ 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 (§ 6.2)	Conduct more research revealing implicit bias through measures like simulated implicit association tests and other psychological frameworks.
	Mitigation, Interpretability and Explainability (§ 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.

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

Table 2: Comparison of the Stereotype Content Model (SCM) and the Agency-Beliefs-Communion (ABC) Model.

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 (1971); Tajfel and Turner (1979); Turner and Reynolds (2003); Ellemers and Haslam (2012)
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 (1987); Ward and Friedman (2006); Gauntlett (2008); Bartlett et al. (2013)
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. (1987); Augoustinos and Walker (1998)
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 (1992); Fiske and Haslam (1996); Fiske and Taylor (2020)
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. (2004); Jost and Van der Toorn (2012); Jost (2019); Banaji (2002)
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.	Wetherell and Potter (1992); Edwards (1991); Augoustinos and Walker (1998)
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.	Crenshaw (2013); Cho et al. (2013); Carastathis (2014)
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. (2002); Cuddy et al. (2011)
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. (2016)
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. (2016)
Five-Tuple Framework	Stereotypes are relational, contextual, and temporally grounded phenomena.	Stereotypes are structured as (Target, Relation, Attributes, Community, Time Interval), enabling computational modeling.	Davani et al. (2025); Shejole and Bhattacharyya (2025)

Table 3: Summary of major theories and frameworks explaining the formation, function, and structure of stereotypes across social psychology and computational social science.

C Summarizing Social-Pyschological Theories and Frameworks

In Section 2, we discussed various theories and frameworks related to stereotypes. We summarize them in Table 3. We compare the SCM Model and the ABC Model in Table 2.

D 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 (Gallegos et al., 2024). 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. (2024) 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 (Cuddy et al., 2008; Fraser et al., 2021), hence this techniques can also be explored for mitigation in the future.

E Use of AI Assistants

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