A Comprehensive Framework to Operationalize Social Stereotypes for Responsible AI Evaluations

Anonymous ACL submission

Abstract

Societal stereotypes are at the center of a myriad of responsible AI interventions targeted at reducing the generation and propagation of potentially harmful outcomes. While these efforts are much needed, they tend to be fragmented and often address different parts of the issue without taking in a unified or holistic approach about social stereotypes and how they impact various parts of the machine learning pipeline. As a result, it fails to capitalize on the underlying mechanisms that are common across different types of stereotypes, and to anchor on particular aspects that are relevant in certain cases. In this paper, we draw on social psychological research, and build on NLP data and methods, to propose a unified framework to operationalize stereotypes in generative AI evaluations. Our framework identifies key components of stereotypes that are crucial in AI evaluation, including the target group, associated attribute, relationship characteristics, perceiving group, and relevant context. We also provide considerations and recommendations for its responsible use.

CONTENT WARNING: This paper contains examples of stereotypes that may be offensive.

1 Introduction & Motivation

Recent years have seen unprecedented gains in generative AI models' capabilities across modalities — language (Anil et al., 2023; Achiam et al., 2023), image (Rombach et al., 2022; Saharia et al., 2022), audio (Kreuk et al., 2022; Borsos et al., 2023), and video (Ho et al., 2022; Bar-Tal et al., 2024), while simultaneously gaining traction in diverse application domains and usage contexts across the globe (Sengar et al., 2024; Raiaan et al., 2024). Along with these advancements, there are growing concerns that these models may reflect, propagate, and amplify societal stereotypes in their predictions and generations (Garg et al., 2018a; Blodgett et al., 2020; Dev et al., 2022; Hovy and Prabhumoye, 2021), and how they may lead to downstream harms (Field et al., 2021; Shelby et al., 2023).

A growing body of empirical work shows how NLP models reflect societal stereotypes about various groups — gender (Bolukbasi et al., 2016), race (Sap et al., 2019), nationality (Jha et al., 2023), disability (Hutchinson et al., 2020), to cite a few. Many of them also build datasets to enable large-scale evaluation of stereotypes in model predictions (Nadeem et al., 2021; Jha et al., 2023; Bhutani et al., 2024). However, current research and resources lack a unified approach towards stereotypes in AI, hindering a comprehensive understanding of the problem space, thereby limiting effective and scalable interventions. First, they fail to capitalize on the underlying common mechanisms that may be contributing to stereotypes in society, data, and models. Additionally, it makes it harder to envision a unified way to tackle and prioritize downstream sociotechnical harms; instead potentially leading to unintended consequences, like new stereotypes emerging when others are mitigated. Another gap stems from adopting simplistic representations of stereotypes for expediency in evaluations, e.g., *(identity, attribute)* pairs overlook core aspects such as how stereotypes tie to specific time and place, which social groups hold certain stereotypes, and what connotations they imply.

Finally, there are different methodologies to source stereotype data — e.g., annotator-driven collection (Nadeem et al., 2021), LLM-enabled collection (Jha et al., 2023), and community centered collection (Dev et al., 2023a), each having unique strengths in terms of scalability, coverage, and reliability. However, we currently do not have an effective approach to determine which of these methods are appropriate in which contexts, what their relative merits (and demerits) are, and how to use these approaches in ways that lean on their strengths and complementarities. Having a unified framework will enable effective intervention, prioritization in high-stake environments, shared knowledge and methods across various efforts to collect data and intervene on models, predictions, and evaluations. Such a framework will also reveal aspects of this problem space that we still have large gaps to fill.

* equal contribution

In order to address these needs, we build off of social scientific theories on stereotypes as well as existing research on evaluating language technologies for stereotypes, and propose a unified comprehensive framework to operationalize stereotype evaluations. Our framework identifies various high level components such as the target group, the attribute associated with the group, the characteristics of their association, the perceiving group, as well as the context within which these stereotypes are prevalent. We also outline a set of recommendations for how to factor in responsibility considerations while using this framework.

2 Background

Social scientists have dedicated substantial research to the study of stereotypes, recognizing their intricate and multifaceted nature (Macrae et al., 1996; Schneider, 2005). This exploration has led to the development of various frameworks over time, aiming to unravel the complexities of how stereotypes originate, function, and influence both individuals and society as a whole (Hilton and Von Hippel, 1996). Early work predominantly viewed stereotypes as inaccurate generalizations about groups, stemming from limited or biased information (Allport et al., 1954). They are also seen as cognitive shortcuts that help individuals simplify and categorize the social world, although this simplification could lead to errors and biases (Dovidio et al., 2010). While these cognitive processes can be efficient, the connection between stereotypes (cognitive bias), prejudice (attitude bias), and discrimination (behavioral bias) was recognized early on, pointing to stereotypes as the motivation for negative attitudes and behaviors towards out-groups (Macrae and Bodenhausen, 2000).

Various theories have been developed that focus on diverse aspects of stereotypes. *Social identity theory* emphasizes the role of group membership in shaping self-concept and inter-group relations, suggesting that stereotypes can serve to enhance one's own group identity (Tajfel et al., 1979). *Social learning theory*, on the other hand, focuses on stereotypes being learned through observation and socialization, often from parents, peers, and media (Bandura and Walters, 1977). *System justification theory* examines how stereotypes can be used to justify existing social hierarchies, even by members of disadvantaged groups (Jost and Banaji, 1994). *Intersectionality theory* further emphasizes the interconnected nature of social identities and how multiple stereotypes can intersect to create unique experiences of discrimination (Crenshaw, 2013).

These theoretical perspectives have guided the development of various frameworks for analyzing stereotypes. Primarily shaped by social psychologists, these frameworks are widely used in other fields to model group dynamics and interactions. One of the prominent such frameworks is the *Stereotype Content Model* (SCM), which posits that stereotypes vary along two dimensions: warmth and competence, resulting in different emotional and behavioral responses towards groups (Cuddy et al., 2007; Fiske et al., 2018). By extending the SCM, the *dual perspectives model* (Abele et al., 2016) added Morality and Sociability axes to the Warmth, and Ability and Assertiveness axes to the Competence dimension. *Agency-Beliefs-Communion* (ABC; Koch et al., 2016) model further added Status to the Competence dimension, and Belief as a dimension; "one end of Beliefs represents all religious, conservative, and other traditional groups; at the other end are progressives, artists, scientists, and LGBTQ groups."

Some of these frameworks are increasingly being explored in NLP research. For instance, SCM has been applied to understand annotator biases (Davani et al., 2023) and debiasing word embeddings (Ungless et al., 2022; Omrani et al., 2023). Fraser et al. (2022) present a computational method to apply SCM to textual data and demonstrated that stereotypes in textual resources compare favourably with survey-based studies in the psychological literature. Fraser et al. (2024) used the ABC dimensions to evaluate and compare biases toward occupational groups across traditional survey-based data and various text sources. As NLP efforts increasingly grapple with the complexities of stereotypes in language, relying solely on social psychological frameworks of stereotypes can limit the scope of the analyses. These frameworks often prioritize dimensions like warmth and competence, potentially overlooking crucial aspects such as social dynamics, socio-historical context, and linguistic valence, which are also essential for a comprehensive understanding of stereotypes in language technologies.

3 Reflective exercise

In this section, we present a reflective exercise on NLP research on social stereotypes with the objective of demonstrating various focus areas surrounding this topic. For comprehensive surveys on this active research area, see Blodgett et al. (2020, 2021).

3.1 Stereotype Detection and Evaluation

A significant number of responsible AI and NLP evaluations are concerned with various concepts that are inherently intertwined with stereotypes. For instance, bias measurement in co-reference resolution tasks often relies on gender-based occupation stereotypes (Zhao et al., 2018; Rudinger et al., 2018), hate speech detection can hinge on societal stereotypes (Chiril et al., 2021), offensive text can be comprised of stereotypes (Jeong et al., 2022), sentiments that are disparately associated with different target groups stem from stereotypical perceptions about them (Kiritchenko and Mohammad, 2018), and more. However, the stereotype resources that these evaluations depend on, are limited in which groups they represent. While substantial work has focused on gender and racial stereotypes, they are also mostly constrained by binary gender constructs (Dev et al., 2021) and Western racial histories (Sambasivan et al., 2021). Other identity axes such as disability status, or socio-economic conditions are not as well represented. These resources are also strife with Western gaze wherein a majority of the resources are collected in the West (or even specifically North America), with data and annotators both representing Western viewpoints.

Based on keyword-based querying of the ACL anthology,[1] we note that 4140 papers mention stereotypes, their detection, resources, and evaluation. Of these, 54.1% mention gender-based stereotypes, 25.8% mention racial stereotypes, and only 16.4% mention region- and nationality-based stereotypes, and an even smaller fraction mention other identities such as age, disability, and profession. Some papers categorize stereotypes as positive or negative, often discussing the associated sentiment rather than the effect it can have downstream, or the specific marginalization the target groups experience (Blodgett et al., 2021). For example, 'women are polite' can arguably be considered positive because of the sentiment associated with politeness, but the stereotype can have other implicit harms (Cheng et al., 2023) related to the history of expectations of politeness and servitude from women (Garg et al., 2018b), something that can negatively influence applications such as job recommendations based on gender.

3.2 Stereotype Resource Creation

Evaluating how stereotypes impact NLP model outputs requires societal data that capture such stereotypes. In this section, we will discuss different approaches used to build such datasets employed in NLP research.

Social psychology studies: Historically, social psychology studies have provided a rich source of societal stereotypes that have been utilized to develop both resources and evaluation strategies for AI models (Caliskan et al., 2017). These studies can contribute societal grounding regarding how a stereotype is perceived (Fiske, 1991; Kite et al., 2022), as well as provide extensive examples of prevalent stereotypes about different groups of people (Borude, 1966) that have been used in NLP evaluations (Bhatt et al., 2022).

Crowdsourcing studies: NLP researchers have recently began adapting social-psychological resources to build NLP evaluation datasets for stereotypes at scale. Approaches such as StereoSet (Nadeem et al., 2021) and CrowsPairs (Nangia et al., 2020) addressed the need for scaling stereotype data via crowdsourcing platforms such as Mechanical Turk. This crowdsourced data, while exceptionally valuable, is often tied to recognizing stereotypes reflected in specific modalities (e.g., recognizing whether a particular text reflects a stereotype), and not as a stand-alone list of social stereotypes as societal knowledge. As a result, the number of identities and unique stereotypes captured in such resources tend to be relatively small.

[1]https://aclanthology.org/

Media crawling: Crowdsourced data, while exceptionally valuable, is often restricted in its media form (primarily text), representation (who participates in crowdsourcing), and time (reflecting a specific moment). Researchers, therefore, turned to "big data" resources (e.g., social networks, and web crawls) which offer a broader range of content, perspectives, and temporal data. Existing media content, whether text, images, or videos, is shown reflect the stereotypes present in the society. Wikipedia, for instance, documents the origins of some well-known stereotypes and describes their provenance. News articles and social media can (un)intentionally propagate stereotypes as expressed by their authors. A popular approach for collecting such stereotypes is to crawl resources and capture co-occurrences of identity terms and attribute words (Sap et al., 2020; Bhatt et al., 2022; Bourgeade et al., 2023).

Model-generation-based studies: While crowdsourcing and social media based curation increase the scale of stereotype resources, they are still limited in coverage of identities and range of associated stereotypes. More recent approaches has looked into leveraging large language models to expand coverage of stereotypes in a rapidly scalable manner and create a resource with broader coverage. When coupled with human annotations, these approached provide validated resources that even significantly overcome selection bias of data creators (Jha et al., 2023; Bhutani et al., 2024). While this expands the state of stereotype resources across identity axes, languages, and cultures, such an approach holds only when models are exposed (via their training data) to such social information in specific languages and about particular identity groups; thus leaving gaps in coverage across the world and many marginalized groups who are not well-represented in the online discourse.

Community-engaged studies: Marginalized communities, who face some of the most severe stereotypes, are often not represented in most resources that are sourced by the previously mentioned methods. Representation is often influenced by how much these communities are written about, who gets to participate as an annotator or crowd worker, and the limits of participation in any of these roles (Birhane et al., 2022). To circumvent these gaps, recent work has engaged with underrepresented and underserved communities in a targeted manner to bridge the gaps in salient stereotype resources (e.g., (Alemany et al., 2022; Dev et al., 2023a; Ación et al., 2023)).

These approaches often offer complementary strengths and weaknesses (Dev et al., 2023b). E.g., social psychological studies and community sourced studies tend to generate relatively smaller resources, but they bring forth richer and nuanced perspectives such as the perceiver group, and the extent of marginalization of the target group, while filling gaps about communities that are underrepresented in existing resources.

3.3 Gaps in Current Approaches

While the variety of approaches for collecting stereotypes do overlap and address some gaps (e.g., scalability and coverage), significant limitations persists across many of the mentioned approaches.

Stereotypes evolve over time: Stereotypes are not static but rather temporally variable. They are influenced by how terms get reclaimed and change in meaning, historical events that lead to a shift in sentiment towards groups of people, and more (e.g., (Garg et al., 2018b)). Yet, most resources capture stereotypes as a snapshot without capturing their evolving nature. For a resource to be operationalizable in bias mitigation or data and model evaluations, temporal grounding is critical. This helps resolve questions regarding factuality (e.g., French kings in 1600s being White is factual and not stereotypical) and misinformation (current Pope is not female, or Asians being associated with COVID 19 post the pandemic (Lin et al., 2022)), identification of offensive slurs or pejorative terms (e.g., the word *protestant* was derogatory in 1500s but is simply a descriptor of religious identity now) and prevalent discriminatory practices (e.g. fraction of women who could vote in the United States before and after the women's suffrage movement (Garg et al., 2018b)).

Silowed Stereotype Evaluations: Stereotypes affect humans and social interactions. With stereotypes reflected in generative models, they consequently impact human-AI interactions with the potential to cause a range of harmful or unpleasant effects. However, evaluations of stereotyping happen predominantly at the model checkpoints rather than at downstream use cases or applications in everyday life. They are also considered as an evaluation pillar of its own without considering the implications on various other representational or allocational harms (Barocas et al., 2017; Shelby et al., 2023).

Lack of Consistent Conceptualization: As discussed by Blodgett et al. (2021) in a thorough assessment of a number NLP measurements of stereotypes, benchmarks do not always rely on solid conceptualizations of stereotypes. Definitions of stereotype often lack critical components such as power dynamics, and consistency in defining social categories. Moreover, even thorough considerations during conceptualization are not guaranteed to be accurately reflected into operationalization. While these gaps are often hard to completely eliminate, it is important to articulate them to further attention on more effective operationalizations.

Perceiver as a missing piece of the puzzle: While stereotypes are born as interactions between social groups, one being the group that is perceiving and one the group that is being perceived, most frameworks and benchmarks do not consider the perceiver group and solely focus on the target group. Notably, Jha et al. (2023) point out that individuals in different geographic regions are familiar with different non-overlapping stereotypes about the same identities. While computational work on stereotypes have expanded the participants pool through crowdsourcing — although the intention for this is often to reduce the cost and time, and not to diversify the sample, they still do not take the crowdworkers background information into account in how these resources are used.

Lack of Contextual/Societal Grounding: Not every over-represented association is a stereotype. Stereotypes require societal grounding for identification of harms caused (Bhatt et al., 2022). Large-scale model evaluations for stereotypical or 'biased' behavior without contextual grounding merely calibrates model tendencies. A common example is racial bias and specifically anti-African-American stereotypes that are prevalent in the United States and rooted in colonial history, but are not as prevalent in South Asia where skin color does not correlate with race or nationality. Grounding a stereotype with what specific socio-cultural settings it is common in, helps build better evaluation paradigms and generative AI systems (Sambasivan et al., 2021).

Multilingual and Multi-Cultural settings Stereotypes are often erroneously considered as absolute, intransient features of society that translate perfectly through languages and cultures. This however has been noted to be objectively incorrect (Cuddy et al., 2009), with distinct stereotypes existing in different geo-cultures (Malik et al., 2022; Bhatt et al., 2022), some of which are expressed with words that are salient in only one language (Bhutani et al., 2024).

4 Framework

Typically, stereotypes generalize certain social groups with specific traits that allude to their agency (*competence*), experience (*warmth*), and often even their *morality*. This is rooted in the underlying cognitive process of categorizing, which helps humans make sense of the world by allowing them to track and distinguish others while using only a small amount of cognitive resources. We build on the social psychological conceptualization of stereotypes to introduce a framework for formalizing and depicting the content of a stereotype. Our framework is composed of five main components: the **target group**, the associated trait or **attribute**, the **association** between the target group and the attribute, the **perceiver** who holds the stereotypical belief, and the **context** in which this stereotype gets its meaning. Figure 1 summarizes this framework. We now describe each of these five components below.

Target Group The cognitive process of categorizing encourages people to think in terms of "us" (in-group) vs. "them," (out-group) which in turn leads to stereotyping. The out-group, or *target group*, is an integral component of a stereotype, and can be characterized with the following features:

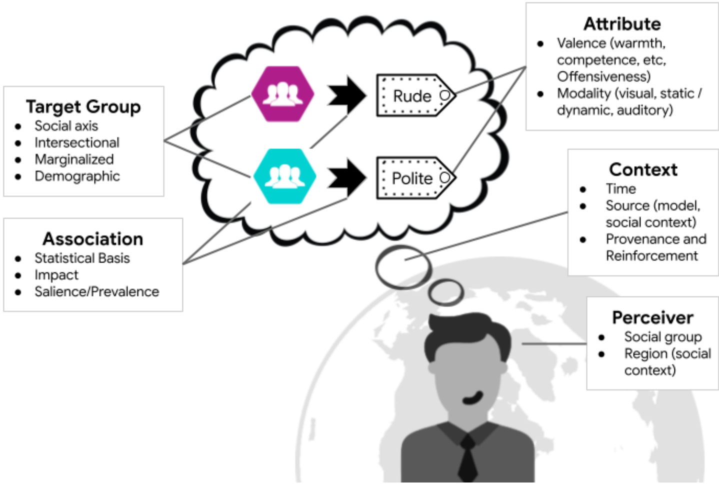

Figure 1: Framework for operationalizing stereotypes.

- *Social axis.* In a social setting what separates individuals from out-groups is their perceived membership in social groups along some social axes (e.g., race, gender, ethnicity). As stereotypes are shaped by societal power structures and historical contexts, understanding the target group's socio-demographic axes helps uncover the factors that contribute to the formation and perpetuation of stereotypes. Not all social groups may be determined in terms of demographic attributes (e.g., one may hold stereotypes about *techies*, or workers in the technology sector, a social group defined in terms of occupation).
- *Intersectional.* Theories of social categorization explain that perceiving an individual as a member of multiple groups (either considered as the perceiver's in-groups or out-groups) leads to specific stereotypes beyond the ones associated to either of the constituent groups. The perceiver's judgement might change when they categorize the target into in-group gender but out-group race, as opposed to out-group gender and race. So whether the group is intersectional or not is an important aspect to capture.
- *Marginalized.* If a social group is historically marginalized, stereotypes portent to result in more harm. This may result in discriminatory hiring practices enabled by AI systems magnifying stereotypes about temperament and suitability of employment about women and African-Americans who are known to have been marginalized in the US (Bertrand and Mullainathan, 2004; Chen, 2023). Capturing such historical marginalization may help determine (and prioritize) the appropriate course of action once stereotypes are detected in model output.
- *Demographic.* A social group can be defined by demographic features such as race, gender, or age, and other extrinsic or acquired attributes such as profession or life style. Non-demographic groups may be more fluid and self-selected, whereas demographic groups are based on fixed or inherent characteristics. Stereotypes about demographic groups are often intertwined with social dynamics and can be associated with systematic discrimination. Therefore, it is important to capture this distinction.

Attribute The *attribute* describes the beliefs, assumptions, features, sentiments, or perceptions that are widely associated with members of the target group. While the association of these attributes to the target group is core to the notion of stereotypes, the attributes themselves can be characterized with certain features:

- *Valence.* The valence of the attribute can include aspects such as the associated perceived offensiveness (Jha et al., 2023), warmth, competence, or morality (Fiske et al., 2018) of the term. The perceptions of attributes as such, and what motivates people to use them, is discussed in social psychology and NLP literature and can inform practices that rely on human ratings for identifying stereotypes. The valence of attributes may also help NLP practitioners prioritize stereotypes (e.g., focusing on stereotypes associating groups with offensive attributes).
- *Modality.* Attributes manifest in different ways across different modalities. For instance, attributes like 'soft spoken' or 'intelligent' can be expressed clearly in text, video, or audio, but less likely to be depicted in images. On the other hand, the markers of 'poverty' can be vastly different in text (e.g., descriptions of poverty) versus image or video (e.g., dusty streets as visual markers of poverty that are not often verbalized). Capturing this nuance is crucial to operationalize such large databases of socio-cultural information into robust model or data interventions.

Association The target group and the associated attribute together constitute the core unit of the *stereotypical association*. The association itself can be characterized by the following features:

- *Statistical Basis (cf. Accuracy).* The distinction between whether an association is a stereotype or factual/definitional is often blurry. For example, while it is true that Hindus often pray in temples, and this association is statistically accurate, generalizing all Hindus as temple-goers can be perceived as stereotyping, as Hinduism (like any religion) in practice encompasses a wide range of rituals beyond temple worship. On the other hand, certain associations may be readily accepted as stereotypes, but also has statistical basis: for instance, some occupational stereotypes found in NLP models align with actual US census data on job distribution (Garg et al., 2018b).

- *Impact*. The impact of associating an attribute to a particular group can be distinct from the attribute's valence in isolation. As such, the same attribute can have varying impacts when associated with different target groups. For example, *dominating* or *bossy* can be seen as slightly offensive, but when stereotypically associated with women, it pertains to professional behavior and competence and can be highly offensive. The impact captures the potential negative result of the association on the target group, distinct from (and orthogonal to) the valence of the attribute.
- *Salience or Prevalence*. The salience or prevalence of the association can be described in various levels. It is useful to distinguish them at least at two levels from an NLP perspective: (1) *model/data/language salience* represents how frequently or prominently the association appears in the model or dataset in a given language and can be measured in different ways (Jha et al., 2023; Bhatt et al., 2022). Model salience can further be an indicator of how likely it is to influence model generations. (2) *Social salience* captures how widespread an association is in society, captured either at a global level, or variations across regions and communities.

Perceiver The stereotypical association of the attribute with the target group is held by a group of people or a section of society, who are the *perceivers* of the stereotype. The socio-economic standing of this group of people, and the fraction of population they account for are some aspects that may be of relevance.

- *Social Group.* The social group that the perceivers belong to is crucial in understanding stereotypes because it significantly influences how they distinguish in-groups from out-groups and consequently perceive and interact with the target group. It is also important to note that any implications of social group membership of perceivers will differ from those of target group's social axes. For instance, whether or not a target group is historically marginalized may be crucial in determining how stereotypes about them may be prioritized in certain contexts, but whether the perceiver group was historically marginalized or not may not be hold the same weightage.
- *Region/Social context.* Social groups often have different levels of power and status in society. This power differential can also influence how stereotypes are formed and perpetuated. Therefore the interaction of perceivers' social group and the target group is meaningful in this context. This dynamic is an important factor for determining the possible harmfulness of the stereotype.

Context Finally, it is crucial to remember that stereotypes are not universal or static. They exist within specific social, cultural, and temporal contexts. Instead of implying that we are speaking about "society" in general, it is important to pinpoint both the time period and the specific reference/artifact (a dataset, a model, a geo-cultural region, etc.) that reflects the societal views in question. This precision will help avoid generalizations and ensures a more accurate analysis of stereotypes.

- *Time*. Stereotypes are dynamic associations, reflecting shifts in social group interactions, cultural norms, and historical events over time. The perceivers' exposure to the evolving information, therefore, alters their existing stereotypes. This is an important aspect to capture in how we operationalize stereotypes in NLP research.
- *Reference*. Stereotypes captured in NLP datasets and models, exist within specific socio-cultural contexts. Their prevalence may vary depending on which slice of society is captured in any specific dataset or model. Hence, it is important to also capture this referential context – i.e., which societal context, and which artifact, whether data or model.
- *Provenance and Reinforcement.* The origin of a stereotype can denote the intent or purpose of reinforcing this belief on a social level. Stereotypes may be rooted in social policies, propaganda, myths or scientific misconceptions. Understanding whether a stereotype originates from scientific, religious, media, or political propaganda maybe helpful for evaluating its social impact.

It is important to also note that the features in the framework may interact with one another. For example, *Christians* are a minority group in India and can be seen as marginalized, whereas, the same group is not similarly marginalized in the US. This difference may influence how stereotypes about the same target group may be dealt with in India vs. the US.

5 Roadmap for Operationalization

The framework presented above is intentionally broad, with the aim to capture all aspects of stereotypes that may be relevant in responsible AI evaluations. There may be crucial considerations that helps when it comes to operationalizing the framework in specific contexts. In this section we provide such a roadmap for implementation and utilizing the framework.

5.1 Recommendations for Implementation

Our framework is conceptual in nature, and is not tied to any particular implementation approach. A simpler implementation, for instance, using spreadsheets or relational databases, may suffice if the evaluation con-

Source	Target Group					Attribute				Perceiver	
							Valence				
	Token	Social Axis	Int.	Marg.	Demo.	Token	Warm.	Compet.	Off.	Social Group	Region
SeeGULL	Palestinian	nationality	F	T	T	aggressive	low	high	high	Middle-eastern	Middle East
	Netherlanders	nationality	F	F	T	blunt	-	high	low	European	Europe
	Afghans	nationality	F	T	T	violent	low	high	high	South-Asian	South Asia
StereoLMs	dentists	profession	F	F	F	weird	-	-	low	-	-
	asians	race	F	F	T	elegant	-	-	low	-	-
	millennials	age	F	F	T	nostalgic	-	-	low	-	-
SPICE	brahmins	caste	F	F	T	vegetarians	-	-	low	Indian	India
	dalits	caste	F	T	T	uneducated	-	low	high	Indian	India
	punjabis	region	F	F	T	fearless	-	high	low	Indian	India
CrowsPairs	old	age	F	F	T	fat	-	-	high	-	US
	native Americans	race	F	T	T	lazy	-	low	high	-	US
	schizophrenia	disability	F	F	F	stupid	-	low	high	-	US
SBF	gay men	SO, gender	T	T	T	disgusting	-	-	high	-	US and Canada
	women	gender	F	F	T	objects	-	low	high	-	US and Canada
	immigrants	nationality	F	T	F	primitive	-	low	high	-	US and Canada

Table 1: The table shows instances of stereotype from five NLP resources – SeeGULL (Jha et al., 2023), Stereotypes in LMs (StereoLMs; Choenni et al., 2021), SPICE (Dev et al., 2023a), CrowsPairs (Nangia et al., 2020), and Social Bias Frames (SBF; Sap et al., 2020) – imported into our framework.

text is narrowly-scoped. Table 1 shows one such tabular form implementation of our framework, where we mapped instances from five stereotype resources that are prominently used in NLP. We chose ∼20 examples from each of the datasets, and mapped the existing information in those datasets onto our framework. This exercise revealed cases where certain features are not applicable (e.g., *vegetarianism* as an attribute does not lend itself to the SCM categories of warmth and competence, as it is based on a religious practice. It also revealed cases where existing datasets lack certain relevant information; e.g., StereoLMs dataset does not capture perceiver information, whereas SeeGULL and SPICE capture regional information of perceivers.

While such a simplistic implementation may suffice for demonstrative purposes, and for small scale evaluations, most real-world scenarios will require a more sophisticated implementation that can account for interrelationships between various elements of the framework. In particular, a knowledge graph based implementation might be especially appropriate in this case, as it will support sophisticated analytics for robust data exploration and visualization, a high level of expressiveness to capture complex contextual and metadata details, adaptability to accommodate evolving insights about stereotypes, and extensibility to incorporate related entities and information from other resources.

Knowledge graphs allow for flexible data modeling (Angles et al., 2017), which is crucial for capturing the evolving nature of stereotypes and their associated attributes. They emphasize relationships, enabling modeling complex relationships (Paulheim, 2017) between stereotypes and other components such as social groups. Knowledge graphs also enable capturing nuanced knowledge about context, such as time, locale, and source provenance associated with stereotypes. Their semantic capabilities enable automated reasoning and insights, with structures suited to complex queries, visualization, and pattern detection (Hogan et al., 2021). Stored in graph databases, knowledge graphs support rapid data retrieval and efficient scaling, aided by query optimization techniques like partitioning and indexing (Angles et al., 2017), making them ideal for downstream mitigation efforts.

5.2 Utilizing the Framework

In this section, we outline some of the ways in which our framework bridges many of of the gaps identified in Section 3.3. Which of those gaps are crucial may vary on a case by case basis. For instance, if an evaluation is aimed to be applied in a mono-lingual monocultural setting, then the geo-cultural specification on stereotypes' context may not be crucial in that case.

Identifying Stereotype Categories: Our framework goes beyond modeling stereotypes as simple relationships between an identity (e.g., Mexicans) and an attribute (e.g., lazy), and enable richer evaluations:

- **Metadata utilization.** One of the highlights of our framework is that it includes metadata that can be used to identify societal stereotypes according to specific criteria. For instance, in addition to being able to extract specific stereotypes (e.g., (*Mexican*, *lazy*), our framework enables us to retrieve categories of stereotypes that are of similar type (e.g., other attributes similar in meaning to *lazy*). This will not only enable robust evaluation, but also identify and efficiently fill gaps in existing resources.
- **Targeted evaluation.** Our framework can facilitate verifying whether model's responses contain specific categories of stereotypes. For instance, one might be interested in stereotypes involving identities related to a particular social axis, such as race, religion, or nationality where the identity might be that of the target group or the perceiver; stereotypes where the target is a marginalized group; stereotypes that are particularly offensive in some context; stereotypes that are prevalent in a particular culture and/or region; and more. A unified framework lends itself for such comprehensive and targeted evaluations.

Assessing Stereotype Evolution: Our framework provides a powerful lens through which we can examine the dynamic nature of stereotypes and their evolution across time and contexts.

- **Temporal evolution analysis.** The temporal dimension in our framework allows us to track how the prevalence, valence, and/or social groups associated with stereotypes have changed over time. For instance, it was shown that gender stereotypes have evolved over time (Garg et al., 2018b), with newer stereotypes emerging in different periods of time. Similarly, on evaluation of stereotypes and associated offensiveness, general trends of perception of different groups of people can be determined.
- **Contextual evolution analysis.** Stereotypes also differ across societal contexts, such as rural versus urban areas, or in different countries and cultures. This contextual evolution analysis can be uniquely conducted with a framework that not only unifies all prevalent stereotype data but also includes additional structured information regarding the perceiver, the marginalization of the target group, and more.

Assessing Perceivers and Context: Beyond simply identifying stereotypes, our framework enables a deeper exploration of how these stereotypes are shaped by and impact different perceivers and social contexts.

- **Differences**. We can analyze stereotypes associated with a particular group according to different perceivers. This might be useful to understand how groups along a given spectrum may perceive a certain relevant group to gauge deeper concerns that perceivers might have about the target. For instance, we could compare the stereotypes held by democrats and republicans in the US towards certain groups of people, such as immigrants, trans people, or atheists.
- **Societal impact.** Stereotypes can have broader implications on society such as discrimination, inequality, or social exclusion. A unified framework enables analyzing impact in a holistic manner, tying to downstream harms (Shelby et al., 2023).
- **Policy impact.** Governance policies can intervene on how technologies attenuate or exacerbate social issues such as stereotypes. Analysis of large scale impact of stereotypes in society can in turn enable impact on policies developed to protect communities and mitigate harms. Additionally, unified stereotype frameworks can enable analyzing the impact of policies on societal change (Curto et al., 2022).

Preventing Silowed Evaluations with Stereotype Interdependency: To fully grasp the complexity of stereotypes, it is crucial to move beyond isolated analyses and consider how different stereotypes interact and influence one another.

- **Co-occurrence analysis.** Stereotypes can frequently co-occur, and magnify different aspects of marginalization, such as stereotypes about race and gender, or social class and ethnicity (Bond et al., 2021). Such patterns reveal important interdependencies that our framework enables us to identify in data and models, which in-turn could lead to preventing harms to intersectional groups.
- **Conflict and Synergy analysis.** Multiple stereotypes can exist in a society such that they conflict or contradict each other, leading to social tensions (e.g., immigrants as both lazy and stealing jobs). Stereotypes may also coexist and thus can reinforce or amplify one another, creating a more harmful impact, for instance, black women being stereotypes as loud and angry, can lead to workplace discrimination (Motro et al., 2021). This framework enables analysis and aggregation of such interdependencies at local and global scales.

Detecting Stereotype Origin and Propagation Understanding how stereotypes emerge and spread is essential for developing effective interventions, and our framework provides the tools for tracing these patterns.

- **Influencer analysis.** Stereotypes originate at different points of time and are propagated differently. Recurring examination of resources and models over time helps identify key individuals, groups, or events that have contributed to the creation and/or evolution of stereotypes. For example, around the time of the COVID-19 outbreak and pandemic, anti-Asian sentiment and stereotypes were on the rise, which has been markedly observed (Lin et al., 2022). Similar analysis can help understand the origin, propagation and severity of stereotypes.
- **Media analysis.** The media often plays a critical role in shaping the perception of people worldwide,[2] and in turn it also captures and reinforces perceptions of people already present in society.[3] Analyzing media representations, such as movies, television shows, or news articles, contribute to understanding of the formation and/or reinforcement of stereotypes.

6 Discussion

Stereotypes are known to be prevalent in everyday societal interactions, which in turn find their way into the massive datasets used to train language models, and are often reflected in those model predictions. While researchers have made efforts to evaluate and mitigate this issue, existing approaches on this topic often lack a cohesive structure and tend to focus on isolated aspects of the problem. In this paper, we introduced a comprehensive framework for systematically evaluating stereotypes in language technologies, focusing on five core aspects: target group, attribute, association, perceiver, and context, each with an associated set of features. We also outline detailed guidelines for implementing and utilizing this framework in practice. By offering a holistic and operationalizable approach, we aim to empower researchers and developers to develop responsible AI approaches towards language technologies that are effective and robust.

[2] https://www.chicano.ucla.edu/files/news/NHMCLatinoDecisionsReport.pdf

[3] https://blog.google/intl/en-in/company-news/using-ai-to-study-demographic-representation-in-indian-tv/

7 Limitations

While our framework captures various aspects of stereotyping by drawing from social psychology and NLP, we acknowledge its potential limitations. First, our goal is for the framework improve stereotype evaluation and mitigation in LLMs. This inherent subjectivity in interpreting the application can limit the generalizability of the framework to other NLP tasks. Second, while out framework emphasizes the essential role of "context" in shaping stereotypes, we recognize that context is inherently multifaceted and dynamic, encompassing a vast array of factors, including but not limited to social norms, historic events, individual experiences, and power dynamics. Due to this complexity, any attempt to model the context is inevitably meant to fail. Instead, we encourage researchers to explicitly consider and document the relevant contextual factors in their efforts, even if the those factors expand beyond the specific elements included in current framework. Ongoing critical engagement and reflection is highly necessary for social and historical grounding of stereotype evaluations.

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*, 7:219720.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Laura Ación, Laura Alonso Alemany, Luciana Benotti, Matías Bordone, Beatriz Busaniche, Lucía González, and Alexia Halvorsen. 2023. A tool to overcome technical barriers for bias assessment in human language technologies (edia paper). *Inteligencia Artificial Feminista: hacia una agenda de investigación para América Latina y el Caribe*.

Laura Alonso Alemany, Luciana Benotti, Hernán Maina, Lucía González, Mariela Rajngewerc, Lautaro Martínez, Jorge Sánchez, Mauro Schilman, Guido Ivetta, Alexia Halvorsen, et al. 2022. A methodology to characterize bias and harmful stereotypes in natural language processing in latin america. *arXiv preprint arXiv:2207.06591*.

Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. 1954. The nature of prejudice.

Renzo Angles, Marcelo Arenas, Pablo Barceló, Aidan Hogan, Juan Reutter, and Domagoj Vrgoč. 2017. Foundations of modern query languages for graph databases. *ACM Computing Surveys (CSUR)*, 50(5):1–40.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

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

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. 2024. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual conference of the special interest group for computing, information and society*, page 1. New York, NY.

Marianne Bertrand and Sendhil Mullainathan. 2004. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Re-contextualizing fairness in NLP: The case of India. 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 727–740, Online only. Association for Computational Linguistics.

Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. SeeGULL multilingual: a dataset of geo-culturally situated stereotypes. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 842–854, Bangkok, Thailand. Association for Computational Linguistics.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? opportunities and challenges for participatory ai. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

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.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Keosha T Bond, Natalie M Leblanc, Porche Williams, Cora-Ann Gabriel, and Ndidiamaka N Amutah-Onukagha. 2021. Race-based sexual stereotypes, gendered racism, and sexual decision making among young black cisgender women. *Health Education & Behavior*, 48(3):295–305.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.

Ramdas Borude. 1966. Linguistic stereotypes and social distance. *Indian Journal of Social Work*, 27(1):75–82.

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.

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

Z. Chen. 2023. 10.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2021. "be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2833–2844, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rochelle Choenni, Ekaterina Shutova, and Robert van Rooij. 2021. Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1477–1491, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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 JC Cuddy, Susan T Fiske, and Peter Glick. 2007. The bias map: behaviors from intergroup affect and stereotypes. *Journal of personality and social psychology*, 92(4):631.

Amy JC Cuddy, Susan T Fiske, Virginia SY Kwan, Peter Glick, Stéphanie Demoulin, Jacques-Philippe Leyens, Michael Harris Bond, Jean-Claude Croizet, Naomi Ellemers, Ed Sleebos, et al. 2009. Stereotype content model across cultures: Towards universal similarities and some differences. *British journal of social psychology*, 48(1):1–33.

Georgina Curto, Nieves Montes, Carles Sierra, Nardine Osman, and Flavio Comim. 2022. A norm optimisation approach to sdgs: tackling poverty by acting on discrimination. In *International Joint Conference on Artificial Intelligence*.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. 2023a. Building socio-culturally inclusive stereotype resources with community engagement. In *Advances in Neural Information Processing Systems*, volume 36, pages 4365–4381. Curran Associates, Inc.

Sunipa Dev, Akshita Jha, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. 2023b. Building stereotype repositories with complementary approaches for scale and depth. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 84–90.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, et al. 2022. On measures of biases and harms in nlp. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267.

John F Dovidio, Miles Hewstone, Peter Glick, and Victoria M Esses. 2010. Prejudice, stereotyping and discrimination: Theoretical and empirical overview. *Prejudice, stereotyping and discrimination*, 12:3–28.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in nlp. 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 1905–1925.

Susan T Fiske. 1991. Social cognition.

Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2018. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Social cognition*, pages 162–214. Routledge.

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, 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.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018a. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018b. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

James L Hilton and William Von Hippel. 1996. Stereotypes. *Annual review of psychology*, 47(1):237–271.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37.

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

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501.

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, Abu Dhabi, United Arab Emirates. 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.

John T Jost and Mahzarin R Banaji. 1994. The role of stereotyping in system-justification and the production of false consciousness. *British journal of social psychology*, 33(1):1–27.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Mary E Kite, Bernard E Whitley Jr, and Lisa S Wagner. 2022. *Psychology of prejudice and discrimination*. Routledge.

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.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*.

Hao Lin, Pradeep Nalluri, Lantian Li, Yifan Sun, and Yongjun Zhang. 2022. Multiplex anti-Asian sentiment before and during the pandemic: Introducing new datasets from Twitter mining. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 16–24, Dublin, Ireland. Association for Computational Linguistics.

C Neil Macrae and Galen V Bodenhausen. 2000. Social cognition: Thinking categorically about others. *Annual review of psychology*, 51(1):93–120.

C Neil Macrae, Charles Stangor, and Miles Hewstone. 1996. *Stereotypes and stereotyping*. Guilford Press.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. Socially aware bias measurements for Hindi language representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1052, Seattle, United States. Association for Computational Linguistics.

Daphna Motro, Jonathan B. Evans, Aleksander P. J. Ellis, and III Lehman Benson. 2021. Race and reactions to women's expressions of anger at work: Examining the effects of the "angry black woman" stereotype.

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.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel 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.

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, Toronto, Canada. Association for Computational Linguistics.

Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508.

Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 315–328.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

David J Schneider. 2005. *The psychology of stereotyping*. Guilford Press.

Sandeep Singh Sengar, Affan Bin Hasan, Sanjay Kumar, and Fiona Carroll. 2024. Generative artificial intelligence: a systematic review and applications. *Multimedia Tools and Applications*, pages 1–40.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 723–741, New York, NY, USA. Association for Computing Machinery.

Henri Tajfel, John C Turner, William G Austin, and Stephen Worchel. 1979. An integrative theory of intergroup conflict. *Organizational identity: A reader*, 56(65):9780203505984–16.

Eddie L Ungless, Amy Rafferty, Hrichika Nag, and Björn Ross. 2022. A robust bias mitigation procedure based on the stereotype content model. *arXiv preprint arXiv:2210.14552*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.