Mirror, Mirror: Exploring Stereotype Presence Among Top-N Recommendations That May Reach Children

ROBIN UNGRUH*, TU Delft, Delft, Netherlands
MURTADHA AL NAHADI*, TU Delft, Delft, Netherlands
MARIA SOLEDAD PERA, TU Delft, Delft, Netherlands

Children form stereotypes by observing stereotypical expressions during childhood, influencing their future beliefs, attitudes, and behavior. These perceptions, often negative, can surface across the many online media platforms that children access, like streaming services and social media. Given that many of the items displayed on these platforms are commonly selected by recommendation algorithms (RAs), it becomes critical to investigate their role in suggesting items that could negatively impact this vulnerable population. We address this concern by conducting an empirical evaluation to gauge the presence of Gender, Race, and Religion stereotypes among the top-10 recommendations generated by a wide range of RAs across two well-known datasets in different domains: Movielens (movies) and GoodReads (books). Results analyses reveal that all RAs frequently recommend stereotypical items. Gender stereotypes are particularly prevalent, appearing in almost every recommendation list and emerging as the most common stereotype. Our results indicate that no algorithm has a consistent tendency towards recommending more stereotypical content; instead, high stereotype presence can be found across recommendation strategies. Outcomes from this work underscore the potential risks that RAs pose to children in perpetuating and reinforcing harmful stereotypes—this prompts reflections on their implications for the design and evaluation of recommender systems.

CCS Concepts: • **Information systems** → **Recommender systems**; • **Social and professional topics** → **Children**; • **General and reference** → *Empirical studies*; • **Applied computing** → Document analysis.

Additional Key Words and Phrases: Children, Stereotypes, Recommender Systems, Text Analysis

1 Introduction

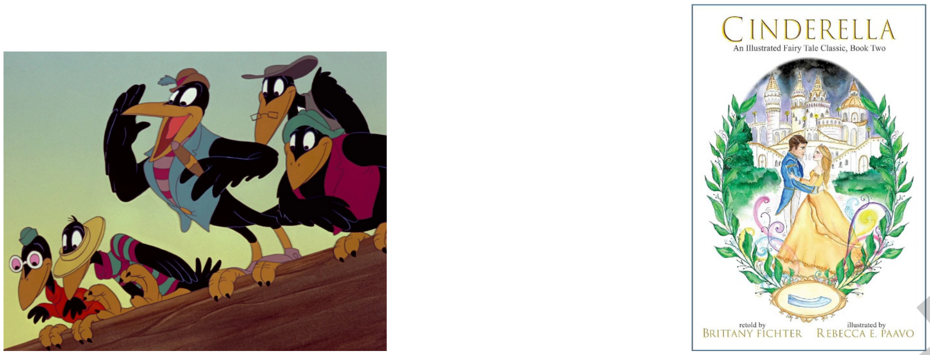
Children are heavily immersed in today's digital world, with online platforms playing an increasingly prominent role in their lives [77, 78]. Not only do they use online platforms to consume entertainment, such as videos and movies, listen to music, or read books, but they can also utilize platforms for educational purposes or be exposed to the marketing of products. Embedded within many e-commerce, social media, and entertainment platforms that children frequently use [78], we find recommendation algorithms (RAs). As a core platform component, RAs select the displayed content by choosing appealing items that match users' preferences or (potential) needs [85]. However, RAs are typically engineered with the 'average adult' user in mind [85], catering recommendations to the preferences of this user group. Considering the unique characteristics of children, it remains uncertain whether recommendations align with children's specific needs, expectations, and preferences, and what type of content children are ultimately exposed to.

RAs can contribute to shaping users' views, preferences, and perceptions in today's online environments [40, 92], making it inevitable that children will be affected by the suggestions these systems make. During

*Both authors contributed equally to this research.

Authors' Contact Information: Robin Ungruh, TU Delft, Delft, Netherlands; e-mail: R.Ungruh@tudelft.nl; Murtadha Al Nahadi, TU Delft, Delft, Netherlands; e-mail: murtadha1997@live.nl; Maria Soledad Pera, TU Delft, Delft, Netherlands; e-mail: M.S.Pera@tudelft.nl.

ACM 2770-6699/2025/3-ART
https://doi.org/10.1145/3721987



(a) A screenshot from Disney's *Dumbo* (1941). ©Disney[1].

(b) Cover of the classical fairytale *Cinderella*, retold by Brittany Fichter and illustrated by Rebecca E. Paavo. ©Brittany Fichter / Rebecca E. Paavo[2].

Fig. 1. Examples of stereotypical depictions of social groups in media.

their developmental phase, when they are actively forming their knowledge and attitudes, children are highly susceptible to the influence of media representations [8, 62]. With media content often serving as a lens through which children perceive the world, in which relatable characters act as role models [7], the presence of harmful content raises concerns about the environments in which children's well-being and safety are shaped.

Among the many harms that can be detrimental to children, a crucial, but often overlooked one lies in perpetuating **stereotypical representations** [83], wherein certain groups are associated with specific characteristics [67]. Children may internalize these representations as realistic depictions of the world around them. Consequently, they may adopt stereotypical beliefs and opinions towards themselves and others [10, 11, 14]. Stereotyping not only results in harmed self-perceptions as well as limited abilities and interests [23, 67, 106], but also in prejudice and negative attitudes against stereotyped groups [10].

Since RAs are a way through which children can discover and access a range of media, the importance of studying the extent to which RAs tend to suggest content that might demonstrate stereotypical attitudes and beliefs becomes apparent. Consider, for example, a recommender on an entertainment platform that includes the items depicted in Figure 1 among its suggestions. Those items are intended for children but are not free from stereotypical expressions. In the movie *Dumbo*, the portrayal of the crows is associated with African-American stereotypes [88]. Similarly, the fairy tale *Cinderella* perpetuates gender stereotypes through its main character's adherence to traditional feminine characteristics [29], a common occurrence in children's literature [101]. Stereotypical representations in children's media have sparked discussions and research about the impact of such stereotypical media on children [6, 21, 105]. While the previously discussed stories intended for children already raise concerns about their impact on children, the fact that recommenders might also expose children to media originally created for adults [48] is alarming.

Regarding the presentation of stereotypical items on online platforms, general discussions on the biases and stereotypes perpetuated by information access systems are of growing interest to the research community [30, 52, 79]. Nonetheless, there is a notable absence of research that specifically addresses how RAs may contribute to exposing children to stereotypical content. This oversight is particularly alarming considering children's formative years. At young ages, children are not just passive consumers of media but are actively engaged,

[1] https://movies.disney.com/dumbo

[2] https://www.brittanyfichterfiction.com/cinderella-an-illustrated-fairy-tale-classic/

impacting the formation of their identities, beliefs, and attitudes in unique ways. The media they are recommended and eventually consume plays a pivotal role in this development. When designing systems that present diverse content, children's needs should not be overlooked. They are distinctly different from adults and are not simply "small adults," but "unique" users [12] who require specific attention.

To shed light on the extent to which RAs can expose children to items conveying varied stereotypes and, hence, might shape children's media experiences and worldviews, we conduct an empirical exploration to examine the tendency of RAs to portray stereotypes among their top-N suggestions. Specifically, we gauge the presence and prominence of Gender, Race, and Religion stereotypes—three common stereotypes [36]—within the top-10 recommendations generated by 14 popular RAs, ranging from naive baseline approaches to sophisticated personalized neural network-based models. Stereotype detection is a complex task [34] without consensus about appropriate strategies to accomplish this automatically; this prompts us to employ 3 models to capture different manifestations of stereotypes. To gain insights into the prevalence of stereotypes across diverse media types, we consider 2 common domains: *Books* and *Movies*. We rely on 2 popular datasets from which we can infer demographic information: Goodreads [104] and MovieLens-1m [44]. To control scope, in this iteration of our work, we determine an item's stereotype based on its *item descriptions*.

With this research, we aim to answer these research questions:

(1) ***RQ1:*** *How prevalent are different types of stereotypes in the textual descriptions of books and movies?*
(2) ***RQ2:*** *To what extent do* RAs *include items that portray stereotypical representations in their top-N suggestions children are exposed to?*
(3) ***RQ3:*** *Is there a trade-off between* RAs*' performance and their tendency to suggest stereotypical items to children?*

To address *RQ1*, we quantify the frequency of stereotypical items in both datasets using different detection models. We analyze and compare the results obtained from these models and discuss the presence of stereotypes across various domains. With respect to *RQ2*, we identify the presence and ranking position of items with stereotypical item descriptions in the top-10 recommendations made to children and compare the tendencies of different RAs to suggest stereotypical items. Finally, we address *RQ3* by identifying differences in the performance of different RAs for children and whether those are linked to the RAs' tendency to suggest stereotypical items.

With this work, we provide an initial assessment of the role that RAs might play in promoting stereotypical content, which we empirically document and analyze. Our empirical setup serves as a blueprint for analyzing harmful content that is recommended to children, which can be reproduced and adapted for further scrutinization of various algorithms, different types of content, and metrics[3].

Lessons learned from our analysis are meant to spark reflection among the research community and beyond regarding the prevalence of stereotypical recommendations to children. This groundwork should prompt further discussions about the severity and potential impact of stereotypical media suggestions but also its potential reasons, leading to a critical assessment of whether such systems can have a detrimental effect on children and how these issues should be addressed.

2 Background & Related Work

In this section, we present background and closely related literature informing our work.

2.1 Stereotypes

Stereotypes refer to beliefs, knowledge, and expectancies that link groups with certain traits or characteristics and are held collectively by a particular society or culture or by individuals [43, 67]. Bigler and Liben [10]

[3] For implementation details and associated code, see https://github.com/Murtadha97/Exploring-Children-s-Exposure-to-Stereotypes-via-Recommender-Systems.git

explore the formation of social stereotypes and prejudice among children, specifically examining how implicit stereotypes are formed unconsciously during childhood. They argue that children categorize individuals by salient dimensions to reduce the complexity of the world. Children internally link traits and attitudes observed in the world toward certain groups. They perceive both the social groups to which they belong and those that differ from them, forming stereotypes through explicit and implicit labeling of these groups [11]. Implicit group categorizations often emerge from how the roles of group members correlate with their real-world representations [10]. For instance, football players are typically associated with male individuals due to their more prominent presentation. Observing and highlighting differences between groups and their attributes often serves as the foundation for stereotypes, resulting in some stereotypes being grounded in actual differences between groups [14]. However, children also form stereotypes that are not necessarily aligned with reality but are learned from explicit expressions and behavior of others [10, 11]. When social groups are labeled, treated, or sorted differently, children tend to categorize them in distinct and meaningful ways. Thus, stereotypes can impact children's personality, values, social identity, and core beliefs [10].

Attitudes, beliefs, and behaviors learned during childhood can persist into adulthood, as this is a critical developmental phase where personality traits like values, social identity, and core beliefs are formed. This is particularly concerning when negative stereotypes are reinforced, as they often lead to prejudice–unfavorable, irrational, and unreasonable attitudes toward the stereotyped group [10]. They may lead to negative beliefs, attitudes, and discriminatory behavior toward others. Typically, children often attribute more positive attitudes towards ingroups than outgroups [10], but self-stereotyping can also negatively influence children's beliefs about themselves and their skills. For example, stereotypes about–among others–race, social class, and gender can create ethnic and racial identities, influencing children's perceptions about their future perspectives and roles [106]. Additionally, children who belong to a gender group generally negatively stereotyped in a STEM field (Science, Technology, Engineering, and Mathematics) tend to doubt their capabilities in this area. Those self-perceptions impact children's development of their own identity [67]. Exposure to stereotypes that evoke negative associations with math skills in the youth can lead to worse achievements in later ages [23]

2.2 Stereotype Detection

Children may encounter stereotypes when interacting with technology, ranging from stereotypical expressions uttered by others on social media [26] to stereotypical content on various platforms [105], which can influence self-perceptions and beliefs about others [11]. Identifying the degree to which children are actually presented with such items often requires computationally detecting stereotypes in media. However, doing so remains challenging and a fairly new research area [34], with approaches mainly stemming from natural language processing and without consent about what approaches actually capture whether an item can be considered stereotypical or not. Stereotype identification approaches use lexicon-based [22], supervised [15, 22, 89] and unsupervised approaches [86]; they also utilize theories from social science [33, 51]. Another recent approach, BiasMeter [36], makes use of the stereotypical tendencies of pre-trained language models to detect stereotypes in texts. One limitation of many proposed methods is that they are often not transparent or generalizable beyond their intended domain or their intended dataset to which they were tuned and tested. This highlights that no unified approach for stereotype detection exists that can serve as a universal ground truth, particularly beyond textual data.

2.3 Stereotypes & Recommender Systems

Recommender systems are often seen as media curators and can leverage word embeddings or other text-based datasets. These sources are prone to perpetuating social biases—such as associating genders with specific professions or personality traits [18, 86]. It is plausible that systems relying on those sources might inevitably replicate such stereotypes, exposing users, including children, to stereotypical representations. Raj and Ekstrand

[82] observe the presence of stereotypes in the ranking of search results for children's products, which reflect Gender stereotypes. This perpetuation of societal expectations through product offerings encourages individuals to continue purchasing items that align with these stereotypical ideas, further reinforcing these stereotypes' prevalence in society. While these findings allude to a tendency for information retrieval methods to propagate stereotypes, there is an absence of research on the connection between RAs and stereotypes.

2.4 Children Recommender Systems

Recommender systems research that considers children often focuses on designing new approaches, mainly for learning or reading materials [57, 60, 69, 93, 103]. These works highlight the opportunities of personalizing content for children in order to suggest content that matches their unique needs, skills, or preferences [38, 70]. That RAs could potentially pose harm to children is frequently discussed [28, 38, 70, 73, 99], but the actual effects of RAs for children and the implications for their well-being are rarely empirically documented [80]. Children are exposed to the decisions of RAs in all kinds of contexts; whether they are the intended user group or not. Thus, the consideration of ethics in designs, particularly towards deploying methods that do not harm children is frequently advocated for [28, 59]. In domains like music [91, 97] children have been considered and analyses have been conducted on classical datasets to explore whether traditional RAs serve children's needs. These investigations, however, are limited to performance or user modeling, i.e., they do not analyze the items generated by RAs regarding potential harms or unsuitable content. Considering the harm that stereotypes could pose to children, there is a need for evaluations explicitly focused on this matter.

3 Experimental Framework

In this work, we study a 'general' recommendation ecosystem, i.e., RAs generate suggestions for their entire user base, including child users who may be exposed to a wide range of content–both intended for them and not. This scenario is quite common, as children often misrepresent their ages on online platforms [65]. Despite age restrictions often placed on users under 13 years old, younger children frequently access these platforms, which allows them to encounter content that may not be directly intended for them [19, 48].

In the rest of this section, we describe the experimental setup for our study. We provide an overview of our experimental pipeline in Figure 2; for implementation details and associated code, see Footnote 3.

3.1 Data

Datasets commonly used for research endeavors related to recommender systems rarely include demographic information, and the presence of children-related samples or info about younger audiences is even rarer.

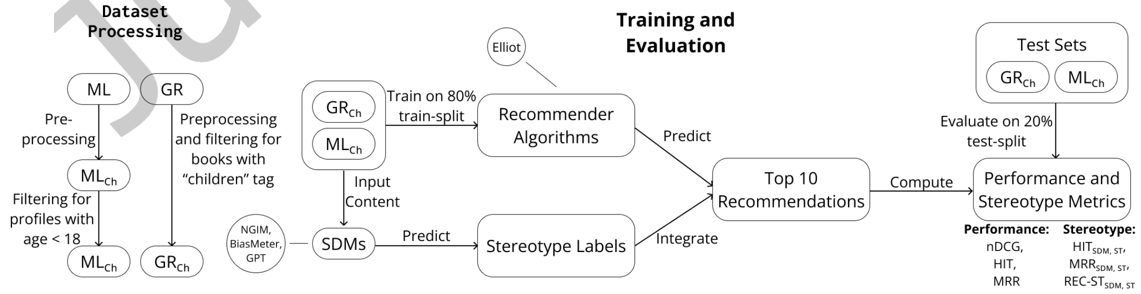


Fig. 2. Overview of the empirical exploration pipeline.

We use two datasets for our exploration: *Goodreads* (GR)[4] [104] and *MovieLens* (ML)[5][44], as these contain real users' preferences for books and movies, are popular among research on recommender systems [85], and include children-related data. **GR** is a large-scale collection of records from users' public bookshelves scraped from the Goodreads website [39, 104], including users' ratings for various items and book metadata like story descriptions. This dataset does not provide users' demographic information. To identify children's profiles which are the focus of exploration, we use a subset of the dataset consisting of books tagged with the "children" genre. We assume the profiles remaining after filtering for such items represent children (or their caretakers) due to the evident interest in children's books. This assumption enables our exploration to gain initial insights in the absence of "real" demographics. **ML** includes entries about users' preferences on a large set of movies in the form of ratings [44]. We specifically use ML-1M, which, to our knowledge, is the only movie-related dataset that includes age information and represents the largest version of the MovieLens datasets that include user demographic metadata, such as age. Since the dataset only provides broad age ranges (Under 18, 18-24, 25-34, 35-44, 45-49, 50-55, 56+) rather than exact ages, we assume that users remain within their assigned age category for all ratings they provide. Thus, we treat all users in the category *Under* 18 as children. We enriched the dataset with item descriptions from *The Movie Database* [100], matching movie titles in ML-1M with descriptions available in the `TMDB 10000 Movies Dataset`[6].

Table 1. Datasets used in our investigation. ML_{Ch} and GR_{Ch} are employed in experiments examining stereotypes in items that could be recommended to children.

Dataset	# users	# items	# interactions
ML	6,040	3,883	1,000,209
ML_{Ch}	5,949	2,654	536,647
ML_{Ch} (testing)	218		
GR	876,145	2,360,655	228,648,342
GR_{Ch}	43,294	14,935	2,483,230

Due to the large size of GR, we focus our exploration on a subset of the item corpus, selecting only books tagged with the "children" genre. This reduces the number of overall user-item interactions and ensures the analysis targets content relevant to children's experiences. In contrast, the ML dataset already contains considerably fewer user-item interactions, so no filtering is applied. In other words, for ML, profiles of adults and children remain in the dataset.

To reduce sparsity, we remove users and items with less than k core interactions in both datasets [4]. We set $k = 10$ for ML following the specifications detailed in [4]. Due to GR's markedly higher sparsity, we set $k = 20$ for GR. This results in GR_{Ch} and ML_{Ch}, respectively. We summarize details on these datasets in Table 1.

3.2 Algorithms

We examine a wide range of RAs from the Elliot framework[7] [5] to offer a comprehensive picture of the type of items different recommendation strategies prioritize. We consider various recommendation generation approaches, ranging from non-personalized naive baselines to personalized and state-of-the-art strategies. For an in-depth discussion on their implementation, along with benchmark performance analysis of these algorithms, refer to [4].

[4] https://mengtingwan.github.io/data/goodreads.html#datasets

[5] https://grouplens.org/datasets/movielens/1m/

[6] https://www.kaggle.com/datasets/muqarrishzaib/tmdb-10000-movies-dataset

[7] https://elliot.readthedocs.io/en/latest/guide/recommenders.html

(I) Naive Baselines:
- Random, a naive baseline, suggests a random selection of items out of the item catalog.
- MostPop, a common alternative, recommends the most popular items based on the number of observed interactions in the training data.

(II) Neighborhood-based:
- ItemkNN is a collaborative filtering approach that suggests items based on similarity to items the user has previously interacted with, using a k-nearest neighbors approach [64].
- UserkNN is a collaborative filtering approach that ranks items based on similarities between the target user and other users, using a k-nearest neighbors approach [87].

(III) Content-based:
- CB is a content-based approach using a vector space model [24] that suggests items to users by calculating cosine similarities between items they have previously interacted with, based on meta-information. In our study, this method is specifically applied to the MovieLens dataset using knowledge graphs sourced from DBpedia[8], which are mapped by Elliot[9] to provide metadata about the items.

(IV) Latent Factor Models:
- MF refers to a "basic" matrix factorization technique that decomposes user-item interactions into latent factors by minimizing the regularized squared error on the set of known ratings [55].
- BPR-MF recommends items by optimizing pairwise rankings using matrix factorization [84].
- FunkSVD recommends items by applying matrix factorization with stochastic gradient descent to uncover latent factors [35].
- PMF recommends items using probabilistic matrix factorization, which models user preferences as probability distributions [72].
- PureSVD recommends items by applying singular value decomposition to the user-item interaction matrix, without regularization [56].
- Slim recommends items by learning sparse linear models that predict item preferences based on past interactions [76].

(V) Artificial Neural Networks:
- DeepFM is a deep learning model that combines convolutional neural networks with factorization machines for feature learning [41].
- NeuMF integrates matrix factorization with neural networks to model user-item interactions [45].

(VI) Autoencoders:
- MultiVAE uses a variational autoencoder that models the probability distribution of user preferences [63].

3.3 Stereotype Detection Models (SDMs)

We focus our exploration on 3 common stereotypes [36]: Gender, Race, and Religion. Currently, there is no universally accepted method for automatically detecting these stereotypes in text [34]. To capture the different manifestations of stereotypes, we employ 3 distinct detection approaches. Utilizing multiple tools allows us to provide a more comprehensive overview of the extent of stereotypes present in the item descriptions of the analyzed content.

Given the scope of our analysis, we focus on detecting stereotypes in textual descriptions, which can provide insights into the items' actual tendencies to promote stereotypical attitudes. Thus, we analyze the item descriptions

[8] https://www.dbpedia.org/

[9] https://github.com/sisinflab/elliot/tree/master/data/cat_dbpedia_movielens_1m_v030

of all items in GR$_{Ch}$ and ML$_{Ch}$[10] with three different SDMs: (i) Name-based Gender Identification Model (NGIM), a naive strategy that offers a simplistic perspective on stereotypes, (ii) BiasMeter, a model based on BERT [36], and (iii) GPT, a model that predicts stereotypes based on responses of the GPT-3.5 model by OpenAI[11].

3.3.1 NGIM. This method identifies Gender inequality based on a simplified two-gender approach, inspired by approaches like the Bechdel test, which famously criticizes the under-representation of female characters in fiction [37]. Various studies show a strong Gender asymmetry and a male bias in media [95, 96]. While not identifying stereotypes directly, this method enables us to gain insights into the over-representation of male characters in both domains. NGIM deems items stereotypical if the count of "male"-represented characters surpasses the number of "female"-represented characters.

Gender stereotypes are identified through gender predictions of unique, named characters featured in the item descriptions of books or the cast of movies, using Python's Gender Guesser library [50]. To retrieve the cast members' names of movies, we use Python's IMDb library [20]. Items for which cast members were not available were treated as non-stereotypical. On book descriptions, Flair's Named Entity Recognition tool [3] extracts characters from the book descriptions.

3.3.2 BiasMeter. Originally developed to identify stereotypes in sentences and documents [36], we adapt BiasMeter to predict the presence of Gender, Race, and Religion stereotypes in the book and movie descriptions. BiasMeter relies on the biased nature of a large language model; it masks words that identify social groups and compares them to the predictions of the language model for those words, producing a score ranging from -1 to 1. Positive values indicate stereotypical tendencies towards the respective groups and negative values indicate anti-stereotypical content.

Since BiasMeter relies on the stereotype tendencies inherent in the language model, its validation against human judgments was demonstrated using individual sentences from two datasets [74, 75]. Thus, we use BiasMeter to score each sentence in a description and aggregate these scores using the Stanford Certainty Factor [66], which is a measure that "reflects how human experts combine and propagate multiple sets of beliefs" [66]. Intuitively, when multiple pieces of evidence are combined, the model adjusts for cumulative certainty: certainty increases when evidence aligns, whereas diminishing returns and conflicting evidence reduce the overall certainty. In our context, this approach aggregates individual sentence stereotype scores, increasing the overall certainty factor as more sentences consistently indicate stereotyping, while accounting for conflicting evidence by balancing positive and negative contributions, so that the final score reflects the overall trend across the entire description. By setting a high-confidence threshold of 0.7, we only classify an item as stereotypical when there is mostly consistent, strong evidence. This aggregation strategy provides a nuanced view, highlighting the presence of stereotypical content in addition to the certainty of its prevalence throughout the description.

3.3.3 GPT. The GPT approach is inspired by Faggioli et al. [31] who show promising results for utilizing large language models for relevance judgments, similar to human ratings. Despite known limitations of LLMs, GPT's power lies in its ability to deal with text-related tasks [46, 58]. Thus, to gain insights and capture aspects the other SDMs may not be able to grasp, such as the interpretation of proper names, which the other approaches fail to do we make requests to the model GPT-3.5 turbo[12] [109] for assessing stereotype probabilities in item descriptions from GR and ML. To do so, we use the following input prompt and feed it to GPT 3.5 in addition to the item description of each item in ML$_{Ch}$ and GR$_{Ch}$[13]:

[10] For a small number of items (< 10%), no item descriptions could be extracted for stereotype detection. Those items were handled as non-stereotypical.

[11] https://platform.openai.com/docs/models/gpt-3-5-turbo

[12] https://platform.openai.com/docs/models/gpt-3-5-turbo, accessed at August, 15th 2024

[13] The variable {type_of_item} is replaced by "books" or "movies", depending on the used dataset.

```
Instruction: You are an expert assessor who is tasked with identifying stereotypes in pieces of text.
The text are descriptions of certain {type_of_item}.
The texts are a few sentences long and if a stereotype is present output the stereotype label and a numbered
     probability estimation of its presence in that text.
The stereotype can have one of the following labels:
    1. Gender
    2. Race
    3. Religion

Please adhere to the following output format and do not include anything else if a stereotype is present:
Stereotype: {stereotype label}, Probability: {the numbered probability as a probability between 0 and 1}
If the stereotype is not present output: 0

Please provide your assessment for the following text:

****Item Description****
```

From the response, items with Gender, Race, or Religion stereotypes yielding a probability of at least 0.7 are labeled as stereotypical.

Approach Validation. To validate the applicability of this approach for stereotype detection, we conducted an analysis similar to the validation of BiasMeter by Gaci et al. [36]. We utilize the CrowS-Pairs [75] and StereoSet [74] datasets as ground truth. These datasets include sentences with human-generated labels indicating their representation of stereotypical expressions. These labeled sentences serve as ground truth, allowing us to evaluate whether our method accurately detects stereotypical representations in items based on these established benchmarks.

In this case, we treat the sentence as an item description in the prompt above. We then evaluate GPT's ability to correctly identify "stereotypical" items (as annotated in the dataset), and not label those that are labeled as "anti-stereotypical" or "neutral". As reported in Table 2, GPT correctly identifies stereotypes in over 75% of the texts on the CrowS-Pairs dataset. However, its accuracy drops to slightly above 50% on Stereoset. Compared to BiasMeter's validation on the same datasets [36], GPT outperforms on CrowS-Pairs, especially with Religion stereotypes, but underperforms on StereoSet, particularly in identifying Religion stereotypes. Consistent with BiasMeter, Gender stereotypes are the most challenging to identify accurately. Gaci et al. [36] suggest this difficulty arises from CrowS-Pairs' frequent use of proper names, which their method struggles to recognize. GPT performs similarly on CrowS-Pairs but is challenged by StereoSet's structure.

The relatively low accuracy of GPT on the two datasets prompts us to a more in-depth analysis, where we probe items that are deemed stereotypical by GPT. More than 70% of gender stereotypical texts and over 90% of religion stereotypes are correctly identified as stereotypical. This high sensitivity to stereotypes, however, also leads to many sentences being misclassified as stereotypical—often when they simply contain a word associated with a social group. Such misclassifications contribute to the lower overall accuracy, resulting in the treatment of many non-stereotypical items as stereotypical. These insights, along with the fact that accuracy scores are comparable to those reported in [36], support our choice of the use of this approach as one of the lenses for the exploration of stereotypes.

Table 2. Accuracy of correctly predicted items per stereotype per dataset.

	All	Gender	Race	Religion
CrowS-Pairs	0.7508	0.6718	0.7636	0.8857
StereoSet	0.5087	0.5176	0.5218	0.3207

3.4 Metrics

To quantify the **stereotype presence** among top-10 recommendations, we adopt three metrics, each labeled with two suffixes: *SDM* indicates the SDM employed for stereotype detection, and *ST* signifies the stereotype, either Gender, Race, or Religion.

- **HIT**$_{SDM,ST}$ is derived from the HIT metric, which typically measures the presence of at least one relevant item in a list. We adopt this metric to assess the presence of *stereotypical* items within recommendations, representing the proportion of lists containing at least one stereotypical item of type ST, as predicted by an SDM.
- **MRR**$_{SDM,ST}$ adapts MRR which typically measures performance by assessing the ranking position of the first relevant item in a list. Our adapted metric captures the average relative position of the first stereotypical item across all recommendation lists. This metric becomes specifically important when evaluating recommendations for children since those tend to click on the search results on top of the lists, particularly the first item, regardless of their relevance [27, 42, 71]. Hence, stereotypical items ranked higher on the list may pose a greater risk to children compared to those ranked lower.
- **REC-ST**$_{SDM,ST}$ is based on the SERP-MS metric [49]. The original metric accounts for the number of results and their position on a search engine results page that contains misinformation. REC-ST$_{SDM,ST}$ (Equation 1) captures the number of items among the top-10 recommendations that convey a stereotype, while taking into account the ranking position of said items, providing insights into the severity of stereotypical items, with higher-ranked items posing more risk to a child.

$$\text{REC-ST}_{SDM,ST}@N = \frac{1}{|U|}\sum_{u\in U}\frac{\sum_{r=1}^{N}(x_{u,r}\cdot(N-r+1))}{\frac{N\cdot(N+1)}{2}} \tag{1}$$

where U is the set of users for whom the top-N recommendations were created and $x_{u,r}$ marks whether the item at position r in the recommendations for user u has a stereotype of type ST, according to an SDM (1 if it has, 0 otherwise).

To assess the **performance** of RAs [4], we compute the following metrics:

- **nDCG** captures the suggested items' relevance and their ranking in the recommendation list by comparing the actual ranking to the optimal ranking where all relevant items were on top of the recommendation list.
- **MAP** indicates the ability of an RA to rank relevant items at a higher position in the top-N recommendations.
- **MRR** evaluates the rank of the first relevant item in the top-N recommendations.

3.5 Experiments

To address our RQ, we undertake 3 experiments.

Exp1 — Stereotype Presence in Source Data. We identify how frequent stereotypes are among items that are eligible to be recommended, i.e., the underlying item corpus. This provides insights into how prominently stereotypical expressions are included in item descriptions. For this, we quantify the proportion of stereotypical items in ML$_{Ch}$ and GR$_{Ch}$, according to the different SDMs. By applying different methods, we can grasp various facets of stereotypical expressions. We compare the frequency of different types of stereotypes (Gender, Race, and Religion) and discuss what aspects are picked up by the different detection methods.

Exp2 — Stereotype Presence in Recommendations. We assess the extent to which each RA suggests stereotypical items in recommendations presented to children. This offers insights into RAs' susceptibility to propagate and possibly amplify stereotypes in the recommendations that have the potential to affect children. By applying different SDMs, stereotype metrics, and a wide range of RAs, we gain a broad picture of RAs' behaviors. We quantify the presence of stereotypes in the top-10 recommendations generated by the RAs presented in Section 3.2. We train and evaluate the algorithms on both ML_{Ch} and GR_{Ch}. We evaluate stereotype presence using the previously mentioned stereotype metrics (see Section 3.4) and compare whether there are significant differences between the presence of Gender, Race, and Religion stereotypes, as well as between different RAs. To analyze this, we perform a two-way ANOVA and compare metric scores of different RAs using paired t-tests with Bonferroni correction ($p < 0.05$). We discuss differences found in stereotype presence between distinct RAs, different SDMs, types of stereotypes, and datasets.

Exp3 — Recommender Performance. We scrutinize the performance of each RA using the performance metrics introduced in Section 3.4 based on the recommendations generated in Exp2. We do this to gain insights into their ability to generate suitable recommendations *for children* but also enable insights into potential relationships or trade-offs to the presence of stereotypes in generated recommendations. We compare the performance of RAs with paired t-tests with Bonferroni correction ($p < 0.05$).

3.6 Implementation

For algorithmic deployment, we use the Elliot framework [5] and the datasets in Table 1. As in the benchmark evaluation of Anelli et al. [4], we converted ratings into implicit feedback by considering ratings above 3 as positive signals. We partitioned both datasets into 80-20 train-test splits. Unlike GR_{Ch} where we assume that each profile is related to children, the child-related samples are limited in ML_{Ch}. As we are interested in children's profiles and experiences in a general recommendation setting, training splits are sampled from the entirety of ML_{Ch}, but our analysis focuses on the testing splits from the 218 children's profiles included in ML_{Ch} (see Table 1). Doing so allows us to gain insights into children's experiences specifically.

The algorithms' hyperparameters are tuned with the automated algorithm for *Tree Parzen Estimators* [9]. We followed the experimental framework in [4] to determine the parameter ranges used for hyperparameter tuning. For algorithms not discussed in [4], we turned to the papers proposing the models to select appropriate parameter ranges for hyperparameter tuning. Note that we apply each algorithm to both datasets, except CB, which was only applied to ML_{Ch}, due to the availability of a knowledge graph for this dataset [5].

4 Results

In this section, we present the results of the experiments outlined in Section 3.

4.1 Exp1 — Stereotype Presence in Source Data

We quantify the presence of stereotypes in the item corpora considered in our study, which include items eligible to be suggested by RAs. For this, we analyze the number of stereotypical items on ML_{Ch} and GR_{Ch} for each stereotype type. As summarized in Table 3, both datasets have a high number of items deemed stereotypical as per different SDMs.

We find similar trends on both datasets: Gender stereotypes are the most frequently recognized by BiasMeter, followed by Race and Religion. Few items are treated as Race stereotypical, ~9% of the items on ML_{Ch} and ~6% on GR_{Ch}. Religion stereotypes are rarely detected, approximately 2% of the items on either dataset. Generally, GPT tends to deem fewer items as stereotypical than the other SDMs. Of note, BiasMeter predicts more than half of the items to have a Gender stereotype on either dataset whereas GPT predicts about 20% to about 11% on ML_{Ch} and GR_{Ch}, respectively. We see similar effects for Race and Religion stereotypes on GR_{Ch}—GPT deems only about half as many items as stereotypical as BiasMeter does. On ML_{Ch}, the proportion of Race and Religion stereotypically deemed items is comparable to the predictions by BiasMeter, with even slightly higher values by GPT.

Scrutiny of stereotypes detected using NGIM shows inconsistent tendencies across datasets. The frequency of Gender stereotypes in ML_{Ch} based on NGIM is similar to the one computed when using BiasMeter; with

Table 3. Proportion of stereotypical items in ML_{Ch} and GR_{Ch} detected as stereotypical by SDMs.

	SDM	Gender	Race	Religion
ML$_{Ch}$	**NGIM**	0.745		
	BiasMeter	0.628	0.093	0.021
	GPT	0.205	0.080	0.035
GR$_{Ch}$	**NGIM**	0.230		
	BiasMeter	0.547	0.062	0.021
	GPT	0.111	0.036	0.009

NGIM predicting ∼10% more items than BiasMeter as stereotypical. This trend does not hold for GR_{Ch}, where BiasMeter suggests that ∼54% of the items portray a Gender stereotype, and NGIM only 23%. This effect could be traced to the manner in which these models identify stereotypes. NGIM focuses on the over-representation of male characters, whereas BiasMeter considers the presentation of a social group in a stereotypical context [36]. These results evince that items in ML_{Ch} often indicate an over-representation of male characters (NGIM)—in line with prior research indicating this trend in movies [95, 96]—and characters also are presented in gender-stereotypical contexts (BiasMeter). On GR_{Ch}, male characters are less frequently over-represented, but characters still appear in gender-stereotypical contexts.

The difference in the detection of the methods can be seen in Table 4a and 4b. As per the Jaccard similarity, there is a higher overlap between NGIM and BiasMeter on both datasets, GPT tends to differ quite strongly from the other methods, with few items being identified as stereotypical by GPT and another technique as well. Based on Cohen's κ, the consistently low inter-rater agreement between NGIM and BiasMeter, as well as between GPT and the other methods, highlights the variability in how stereotypical representations are identified. This suggests that each method captures a different facet of stereotypicality, and the range of results emphasizes the importance of considering multiple techniques to capture the full spectrum of stereotypical representations.

The salient variability in stereotype detection across SDMs directly impacts our results, as they influence the overall detection of stereotypical content, leading to a high proportion of items being flagged as stereotypical. In ML, 92% of items are identified as having a Gender stereotype by at least one of the SDMs, 16% as having a Race stereotype, and 5% as having a Religion stereotype. For GR, 64% of items are flagged for Gender stereotypes, 9% for Race, and 3% for Religion stereotypes. These results highlight how the choice of detection method can markedly determine whether a prevalence of stereotypes in items is perceived. Relying on a single method may lead to an incomplete or skewed understanding of stereotypical content across different domains.

To exemplify what items are deemed stereotypical by the different SDMs, we present two sample item descriptions from GR_{Ch} below and discuss the results of stereotype detection on these items.

> **Example 1:**
> From the dazzling bestselling duo Jane O'Connor and Robin Preiss Glasser comes a fancy, frilly ballet story with a lot of heart. Young ballerinas and Fancy Nancy fans will shout encore! Fancy Nancy is ready for the spotlight! Fancy Nancy and her best friend, Bree, couldn't be more excited about their upcoming dance show. After all, it's all about mermaids, and who knows how to be a fancy, glamorous mermaid better than Fancy Nancy herself? But when another ballerina wins the coveted role of the mermaid, Nancy is stuck playing a dreary, dull tree. Can Nancy bring fancy flair to her role, even though it isn't the one she wanted? And when disaster strikes right before the big ballet, who will step into the spotlight?Perfect for fans of the Eloise and Olivia books.
>
> – *"Fancy Nancy and the Mermaid Ballet"*
> by Jane O'Connor and
> Robin Preiss Glasser

> **Example 2:**
> As a young black man in the segregated South of the 1920s, Wright was hungry to explore new worlds through books, but was forbidden from borrowing them from the library. This touching account tells of his love of reading, and how his unwavering perseverance, along with the help of a co-worker, came together to make Richard's dream a reality. An inspirational story for children of all backgrounds, Richard Wright and the Library Card shares a poignant turning point in the life of a young man who became one of this country's most brilliant writers, the author of Native Son and Black Boy. This book is the third in a series of biographies by William Miller, including Zora Hurston and the Chinaberry Tree and Frederick Douglass: The Last Day of Slavery. All focus on important moments in the lives of these prominent African Americans.
>
> – *"Richard Wright and the Library Card"*
> by William Miller

Example 1 — "Fancy Nancy and the Mermaid Ballet". This item is deemed Gender stereotypical by BiasMeter and GPT. BiasMeter recognizes that the textual descriptions are stereotypically related to female actors. The biased language model inserts words associated with female actors in positions that describe gender roles (e.g., using words like "she"). When the predicted words align with the gender roles present in the actual text, BiasMeter classifies the description as stereotypical. GPT is asked to provide a probability for the item to be stereotypical. Considering its output, GPT recognizes stereotypical representations and presents a high probability for the book to be actually stereotypical. Interestingly, we find that NGIM does not deem the item to be Gender stereotypical. This is due to the high number of female characters in the book description. NGIM does not find an over-representation of male characters, but GPT and BiasMeter highlight that the female characters are presented stereotypically.

Example 2 — "Richard Wright and the Library Card". The item description of this book mentions a social group directly, i.e. a "black man". GPT assumes a high probability for stereotypical representations of race in this book. BiasMeter, on the other hand, does not recognize that the actors in the description are presented stereotypically. This highlights that instead of stereotypical representations in the description directly, GPT might focus more on the probability that the item represents stereotypes. Considering that the book thematizes segregation and racism against African Americans, stereotypical representations are highly likely, despite the item description not being directly stereotypical.

These examples underscore the importance of employing diverse methods for stereotype detection. How stereotypes manifest in an item can vary greatly, each with different implications for their impact. The three distinct SDMs are designed to capture these varied manifestations of stereotype presence, ensuring a more comprehensive analysis.

4.2 Exp2 — Stereotype Presence in Recommendations

In our analysis of Experiment 2, we first focus on stereotype presence in the top-10 recommendations of ML_{Ch}. We then juxtapose these findings with those from GR_{Ch} and highlight salient differences in observed trends. We summarize the results of our exploration in Table 5.

Using **BiasMeter** for stereotype detection, it is evident that Gender stereotypes are most prominent: $HIT_{BiasMeter,Gender}$=1 for all RAs. This shows that at least one item in each generated recommendation list conveys Gender stereotypes in its descriptions. Gender prominence is further showcased by $\mathrm{MRR}_{BiasMeter,Gender}$ scores that provide insights

Table 4. Comparison of label predictions by different SDMs —NGIM, BiasMeter (BM), and GPT—on both datasets under study.

(a) Jaccard-based similarity.

	$\mathbf{ML}_{Ch}$							$\mathbf{GR}_{Ch}$						
	Gender			**Race**		**Religion**		**Gender**			**Race**		**Religion**	
	NGIM	BM	GPT	BM	GPT	BM	GPT	NGIM	BM	GPT	BM	GPT	BM	GPT
NGIM	1.00	0.53	0.15					1.00	0.27	0.06				
BM	0.53	1.00	0.19	1.00	0.09	1.00	0.20	0.27	1.00	0.11	1.00	0.09	1.00	0.06
GPT	0.15	0.19	1.00	0.09	1.00	0.20	1.00	0.06	0.11	1.00	0.09	1.00	0.06	1.00

(b) Inter-rater agreement.

	$\mathbf{ML}_{Ch}$							$\mathbf{GR}_{Ch}$						
	Gender			**Race**		**Religion**		**Gender**			**Race**		**Religion**	
	NGIM	BM	GPT	BM	GPT	BM	GPT	NGIM	BM	GPT	BM	GPT	BM	GPT
NGIM	1.00	0.06	-0.08					1.00	0.16	-0.04				
BM	-0.08	1.00	0.02	1.00	0.09	1.00	0.31	0.06	1.00	0.04	1.00	0.12	1.00	0.11
GPT	-0.08	0.02	1.00	0.09	1.00	0.31	1.00	-0.04	0.04	1.00	0.12	1.00	0.11	1.00

into the average ranking position of the first stereotypical item, i.e., 2.5 to 3.5, depending on the RA considered. REC-ST$_{BiasMeter,Gender}$ scores ranging from 0.59 to 0.67 highlight a high quantity of Gender stereotypical items in the top-10 recommendations, which are also highly ranked.

Race and Religion, despite being less present than Gender stereotypes, are still found in the recommendations. HIT$_{BiasMeter,Race}$ scores exceed 0.5 for most RAs, suggesting that Race stereotypes can be found in the top-10 recommendations for more than 50% of the users. Religion stereotypes are even less common according to HIT$_{BiasMeter,Religion}$ but still yield scores above 0.1 for all RAs. Furthermore, MRR$_{BiasMeter,Race}$ and MRR$_{BiasMeter,Religion}$ are lower as well, positioning items associated with these stereotypes between below the 8^{th} ranking position (with CB being an exception for both). REC-ST$_{BiasMeter,Race}$ and REC-ST$_{BiasMeter,Religion}$ scores range from 0.053 to 0.092 and 0.012 to 0.021, respectively.

Scrutinizing the predictions by **NGIM**, we find that HIT$_{NGIM,Gender}$ scores are similar to those of HIT$_{BiasMeter,Gender}$. However, MRR$_{SDM,Gender}$ and REC-ST$_{SDM,Gender}$ tend to be lower when employing NGIM as opposed to BiasMeter. Comparing BiasMeter to **GPT** shows that the latter predicts fewer Gender and Race stereotypes in recommendations for all RAs comparing respective metrics. However, regarding Religion stereotypes, BiasMeter and GPT yield similar scores in all three metrics. Overall, for GPT and BiasMeter, the same general trend can be seen. Gender stereotypes are the most prevalent, followed by Race and Religion stereotypes.

When shifting our analysis to GR$_{Ch}$, most tendencies observed on ML$_{Ch}$ persist. Gender stereotypes are most frequently included in recommendations, followed by Race and Religion. Of note, a salient difference refers to NGIM's deviation from BiasMeter. In contrast to findings on ML$_{Ch}$, NGIM detects a lower stereotype presence in the recommendations on GR$_{Ch}$ than BiasMeter according to all 3 metrics. The overall scores remain high with HIT$_{NGIM,Gender}$ scores above 0.9 for all RAs, indicating Gender stereotypical items in almost every recommendation list. GPT leads to lower Gender stereotype scores than the other two SDMs on GR$_{Ch}$, in accordance with findings on ML$_{Ch}$. Additionally, GPT predicts fewer Religion stereotypes than BiasMeter in recommendations.

We find statistically significant differences between Gender, Race, and Religion stereotype metrics in the recommendations across recommendation lists created by different RAs. In addition, we probe statistically significant differences in the tendency of RAs in recommending stereotypical items. For brevity, we do not include a comprehensive overview of all tests. Instead, we provide an in-depth discussion centered on HIT$_{BiasMeter,ST}$ on ML$_{Ch}$ as discoveries emerging from analysis based on this metric depict trends found across datasets, different SDMs, and other stereotype metrics. We perform a two-way ANOVA to measure the effect of different RAs and

Table 5. Performance and Stereotype detection metrics using the three different SDMs computed on the top $N = 10$ recommendations on ML$_{Ch}$ GR$_{Ch}$. The highest score on a metric for a personalized RA (excluding Random and MostPop) is bolded if it is significantly higher than the second-highest score.

		Metric	Random	MostPop	ItemKNN	UserKNN	CB	BPRMF	FunkSVD	MF	PMF	PureSVD	Slim	DeepFM	NeuMF	MultiVAE
MovieLens (ML$_{Ch}$)	**Perf.**	nDCG	0.004	0.122	0.244	0.259	0.122	0.248	0.201	0.187	0.239	0.235	0.269	0.251	0.239	0.247
		MRR	0.009	0.227	0.467	0.481	0.281	0.448	0.372	0.365	0.434	0.440	0.477	0.437	0.438	0.435
		MAP	0.003	0.118	0.233	0.250	0.119	0.220	0.194	0.180	0.225	0.229	0.259	0.229	0.227	0.228
	Gender	HIT$_{NGIM}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		MRR$_{NGIM}$	0.872	0.647	0.913	0.877	0.941	0.858	0.916	0.913	0.854	0.979	0.932	0.875	0.882	0.925
		REC-ST$_{NGIM}$	0.754	0.817	0.873	0.847	0.853	0.826	0.863	0.871	0.841	0.926	0.890	0.855	0.833	0.857
		HIT$_{BiasMeter}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		MRR$_{BiasMeter}$	0.788	0.835	0.792	0.810	0.830	0.783	0.804	0.819	0.829	0.778	0.748	0.813	0.764	0.782
		REC-ST$_{BiasMeter}$	0.619	0.589	0.673	0.641	0.656	0.642	0.640	0.636	0.651	0.648	0.647	0.641	0.623	0.659
		HIT$_{GPT}$	0.881	0.996	0.849	0.904	0.909	0.936	0.812	0.877	0.890	0.872	0.881	0.895	0.867	0.831
		MRR$_{GPT}$	0.394	0.955	0.417	0.502	0.413	0.531	0.374	0.403	0.483	0.381	0.443	0.466	0.372	0.364
		REC-ST$_{GPT}$	0.193	0.428	0.192	0.220	0.202	0.241	0.187	0.191	0.216	0.171	0.199	0.228	0.187	0.188
	Race	HIT$_{BiasMeter}$	0.661	0.220	0.546	0.528	0.486	0.500	0.564	0.546	0.532	0.624	0.541	0.550	0.578	0.628
		MRR$_{BiasMeter}$	0.206	0.028	0.164	0.138	0.109	0.151	0.168	0.181	0.138	0.168	0.147	0.163	0.211	0.233
		REC-ST$_{BiasMeter}$	0.084	0.011	0.069	0.061	0.053	0.062	0.071	0.073	0.060	0.071	0.071	0.066	0.089	0.092
		HIT$_{GPT}$	0.478	0.294	0.253	0.207	0.326	0.253	0.322	0.271	0.289	0.326	0.285	0.257	0.335	0.340
		MRR$_{GPT}$	0.143	0.038	0.071	0.059	0.083	0.069	0.092	0.085	0.081	0.077	0.082	0.066	0.080	0.091
		REC-ST$_{GPT}$	0.066	0.015	0.028	0.023	0.036	0.027	0.036	0.033	0.032	0.032	0.032	0.026	0.032	0.038
	Religion	HIT$_{BiasMeter}$	0.174	0.835	0.142	0.170	**0.486**	0.174	0.151	0.179	0.202	0.165	0.138	0.147	0.161	0.197
		MRR$_{BiasMeter}$	0.068	0.116	0.032	0.041	**0.207**	0.039	0.035	0.050	0.047	0.051	0.031	0.036	0.039	0.052
		REC-ST$_{BiasMeter}$	0.021	0.048	0.012	0.015	**0.068**	0.017	0.014	0.020	0.019	0.018	0.012	0.015	0.016	0.021
		HIT$_{GPT}$	0.230	0.000	0.124	0.092	0.143	0.092	0.138	0.193	0.106	0.092	0.124	0.097	0.166	0.184
		MRR$_{GPT}$	0.103	0.000	0.028	0.018	0.032	0.021	0.038	0.050	0.022	0.022	0.025	0.026	0.036	0.045
		REC-ST$_{GPT}$	0.033	0.000	0.013	0.007	0.014	0.010	0.015	0.018	0.009	0.008	0.010	0.011	0.014	0.018
Goodreads (GR$_{Ch}$)	**Perf.**	nDCG	0.001	0.151	0.331	0.329		0.190	0.185	0.162	0.170	0.241	0.342	0.227	0.244	0.320
		MRR	0.002	0.316	0.567	0.577		0.377	0.336	0.299	0.318	0.458	0.589	0.403	0.435	0.547
		MAP	0.001	0.140	0.312	0.311		0.177	0.165	0.144	0.145	0.220	0.324	0.203	0.223	0.296
	Gender	HIT$_{NGIM}$	0.928	0.997	0.930	0.955		0.988	0.960	0.953	0.984	0.952	0.939	0.938	0.944	0.953
		MRR$_{NGIM}$	0.432	0.764	0.484	0.531		0.623	0.508	0.535	**0.787**	0.536	0.451	0.545	0.518	0.504
		REC-ST$_{NGIM}$	0.231	0.366	0.282	0.315		**0.328**	0.290	0.303	0.323	0.281	0.275	0.297	0.287	0.301
		HIT$_{BiasMeter}$	1.000	1.000	0.996	0.998		1.000	0.999	0.999	1.000	1.000	0.997	0.997	0.997	0.998
		MRR$_{BiasMeter}$	0.729	0.998	0.759	0.834		0.929	0.808	0.794	**0.973**	0.860	0.767	0.813	0.788	0.806
		REC-ST$_{BiasMeter}$	0.547	0.860	0.621	0.708		0.776	0.662	0.637	**0.834**	0.743	0.648	0.656	0.647	0.674
		HIT$_{GPT}$	0.667	0.940	0.495	0.589		0.601	0.542	0.598	**0.814**	0.476	0.508	0.624	0.543	0.519
		MRR$_{GPT}$	0.242	0.155	0.187	0.208		0.179	0.174	0.207	**0.266**	0.206	0.186	0.218	0.204	0.188
		REC-ST$_{GPT}$	0.104	0.099	0.083	0.094		0.083	0.077	0.088	**0.134**	0.071	0.077	0.102	0.094	0.084
	Race	HIT$_{BiasMeter}$	0.474	0.644	0.500	0.563		0.615	0.547	0.475	**0.644**	0.528	0.488	0.566	0.525	0.531
		MRR$_{BiasMeter}$	0.159	0.453	0.172	0.217		**0.310**	0.202	0.160	0.190	0.183	0.144	0.226	0.184	0.168
		REC-ST$_{BiasMeter}$	0.063	0.110	0.068	0.079		0.090	0.074	0.061	0.083	0.065	0.061	0.082	0.071	0.069
		HIT$_{GPT}$	0.289	0.004	0.158	0.147		0.138	**0.242**	0.163	0.148	0.203	0.161	0.208	0.151	0.188
		MRR$_{GPT}$	0.091	0.001	0.042	0.043		0.024	**0.075**	0.048	0.031	0.056	0.048	0.061	0.041	0.057
		REC-ST$_{GPT}$	0.034	0.001	0.017	0.016		0.011	**0.028**	0.018	0.014	0.020	0.017	0.024	0.017	0.022
	Religion	HIT$_{BiasMeter}$	0.187	0.035	0.224	0.178		0.189	0.222	0.222	0.078	**0.289**	0.264	0.193	0.216	0.199
		MRR$_{BiasMeter}$	0.057	0.004	0.071	0.054		0.032	0.062	0.059	0.016	0.080	0.082	0.058	0.065	0.064
		REC-ST$_{BiasMeter}$	0.021	0.001	0.027	0.020		0.013	0.023	0.023	0.006	0.032	0.031	0.023	0.025	0.023
		HIT$_{GPT}$	0.083	0.000	0.018	0.040		0.010	**0.125**	0.114	0.027	0.012	0.030	0.054	0.022	0.037
		MRR$_{GPT}$	0.026	0.000	0.005	0.013		0.002	**0.047**	0.035	0.007	0.002	0.008	0.015	0.005	0.010
		REC-ST$_{GPT}$	0.009	0.000	0.002	0.005		0.001	**0.015**	0.012	0.003	0.001	0.003	0.005	0.002	0.004

types of stereotypes (Gender, Race, or Religion) on $\text{HIT}_{BiasMeter,ST}$ scores. We find statistically significant main effects of the RAs ($F(11.64, 2526.62) = 9.117$, $p < .001$, $\eta^2 = 0.010$) and stereotypes ($F(1.68, 363.67) = 1083.002$, $p < .001$, $\eta^2 = 0.443$) as well as a significant interaction effect ($F(17.04, 3697.29) = 44.502$, $p < .001$, $\eta^2 = 0.094$). We highlight the main effect of RAs, as we can observe that Gender stereotypes have the highest $\text{HIT}_{BiasMeter,ST}$ ($M = 1.000$, $SD = 0.000$) followed by Race ($M = 0.542$, $SD = 0.498$) and Religion ($M = 0.227$, $SD = 0.419$). Pairwise t-tests highlight that these differences are statistically significant $p < 0.05$. This analysis reveals disparities in the prevalence of 3 types of stereotypes and the varying tendencies of RAs to expose children to stereotypical content.

To scrutinize these significant differences between RAs found by the two-way ANOVA and observe whether some RAs are consistently more prone to recommending stereotypical items to children, we conduct paired t-tests with Bonferroni correction, comparing the stereotype metric scores across pairs of RAs (see Table 3). We note significant differences between most pairs of RAs across stereotype metrics on GR_{Ch}. We can observe that most cells are colored red, indicating statistically significant differences. However, no RA has a consistent trend towards a higher or lower presence of stereotypes in its recommendation lists across metrics and stereotype types. For example, while FunkSVD stands out on GR_{Ch} in presenting the highest number of Race and Religion stereotypes as predicted by GPT, this trend cannot be observed for other metrics or when using other SDMs to measure stereotypes. Similarly, PMF stands out on GR_{Ch} as frequently including most Gender stereotypes in its recommendations, as predicted by various metrics—a trend that does not translate to other stereotypes.

In contrast, on ML_{Ch}, there are fewer statistically significant differences between RAs. The only personalized algorithm that stands out is CB. On several metrics, it emerges as the sole personalized algorithm exhibiting several statistically significant differences compared to other RAs. Take the metrics related to Religion stereotypes identified by BiasMeter. According to $\text{HIT}_{BiasMeter,Religion}$, Religion stereotypes appear more than twice as frequent as for any other personalized RA. Similar effects are found for $\text{MRR}_{GPT,Gender}$ and $\text{REC-ST}_{GPT,Gender}$. Our examination of personalized RAs indicates that only the content-based algorithm demonstrates increased stereotype exposure across multiple metrics compared to other personalized RAs. However, this effect is specific to ML_{Ch}, as this algorithm was exclusively applied to this dataset.

MostPop's tendency to recommend stereotypical items differs from personalized RAs considered in our study across both ML_{Ch} and GR_{Ch} datasets. Stereotype metrics scores for Race and Religion stereotypes computed on recommendations generated by this widely used strategy are significantly different from those resulting from most personalized counterparts on ML_{Ch}. However, it is not possible to assert whether MostPop is responsible for higher or lower stereotypes exposure in its recommendations in comparison to personalized counterparts. Instead, the impact of MostPop on stereotype exposure depends on the specific metric being considered. $\text{REC-ST}_{BiasMeter,Gender}$ is lower for MostPop compared to all personalized RAs except MF and NeuMF. Using BiasMeter for stereotype prediction on ML_{Ch}, Race stereotypes are less frequent in MostPop's recommendations based on all metrics. However, Religion stereotypes are more common based on all metrics using BiasMeter. Since MostPop inherently prioritizes the most popular, it becomes evident that these items greatly vary in their stereotypical content depending on the stereotype and the SDMs used. On GR_{Ch}, in contrast, the trends found for MostPop are more uniform: Gender stereotypes are found significantly more frequently included and ranked higher in MostPop's recommendations compared to all other lists, as indicated by all metrics using all SDMs. Race stereotypes are more frequent in its recommendations than those of any other algorithm according to BiasMeter. Yet, Religion stereotypes are less frequent in MostPop's recommendations based on this SDM. GPT assumes very few Race and Religion stereotypes in MostPop's recommendations.

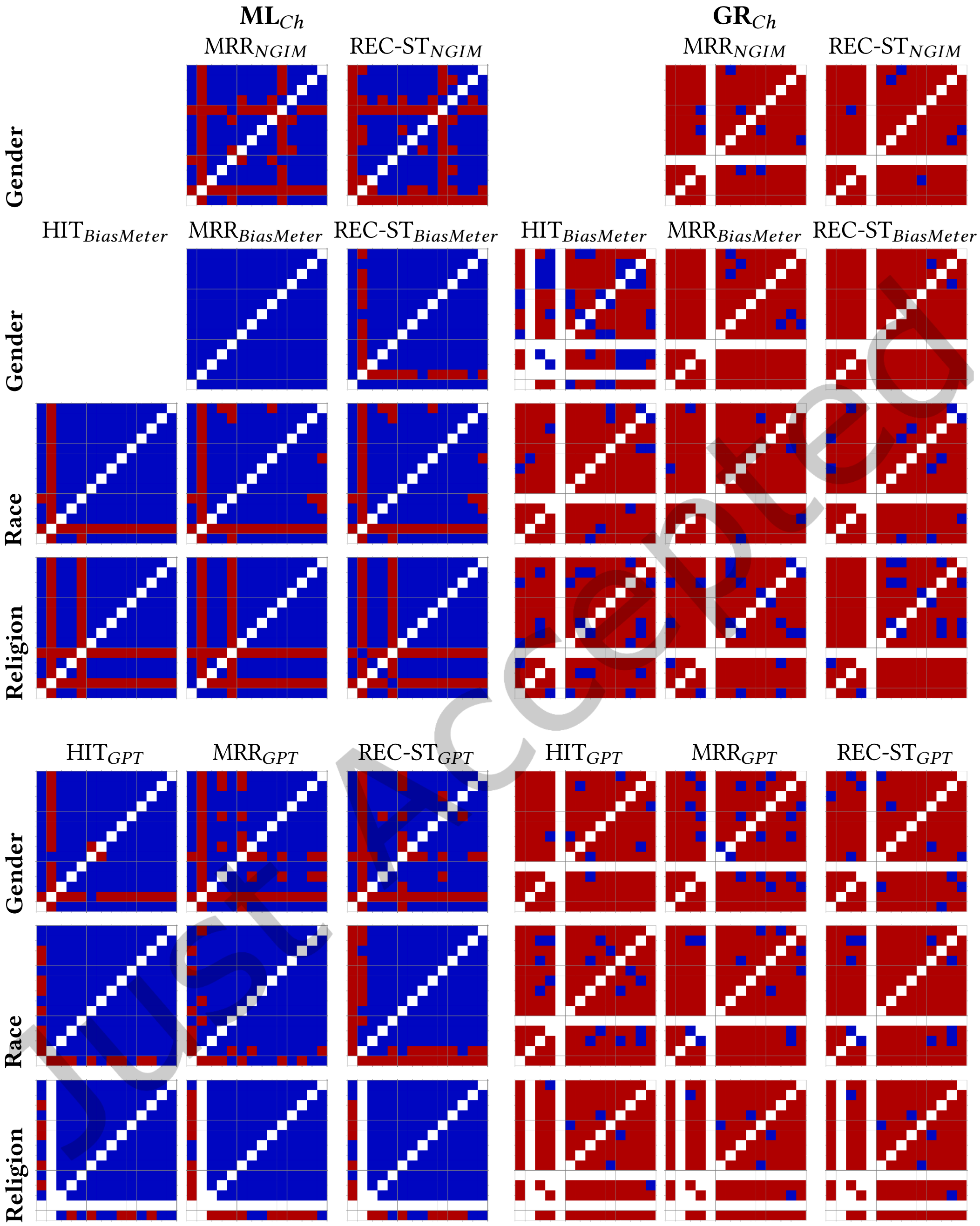


Fig. 3. Paired t-tests with Bonferroni correction ($N = 91$ for ML_{Ch}, $N = 78$ for GR_{Ch}) on stereotype detection metrics computed on recommendations created by different RAs. $p < .05$ indicated by red cell color. Order from left to right and from bottom to top equal to Table 5. Note that for some pairs of RAs for certain metrics no tests could be computed since metric scores were too similar. Those cells are colored white. Since CB was not applied to GR_{Ch}, related cells remain white.

4.3 Exp3 — Recommender Performance

To offer context on how RAs fare, regarding offering children relevant suggestions, we examine their effectiveness using varied metrics, which we report in Table 5.

The results on ML_{Ch} reveal a noticeable decrease in performance across all metrics for all RAs when compared to well-known benchmarks on the original ML dataset evaluated using Elliott [5]. As our study assesses algorithmic effectiveness on children's profiles, and not all the profiles in ML, deviations from benchmarks are anticipated. Still, performance trends remain similar to the reported benchmarks [4]. We rank the algorithms using their scores based on nDCG@10 on ML_{Ch}. The Kendall's τ correlation coefficient, measuring the similarity between algorithm rankings, is 0.7857 when comparing the rankings reported in [4] with those from our analysis, which specifically focuses on children's preferences. This high value indicates a similar ranking, suggesting that the relative effectiveness of algorithms remains consistent when focusing specifically on children compared to the broader evaluation in prior work. Regardless of the performance metrics considered, we observe similar scores for most personalized RAs. On ML_{Ch}, except for CB, nDCG scores range from 0.187 to 0.269, indicating a relative lack of relevant items positioned high in the ranking among top-10 suggestions. MRR scores range from 0.365 to 0.481, highlighting that the first relevant item often appears in the bottom half of recommendation lists; MAP scores between 0.187 and 0.259 indicate that, on average, only a small proportion of relevant items are positioned near the top of recommendation lists. With CB and the baseline RAs, Random and MostPop being exceptions, only a few RAs perform significantly better or worse than others, according to paired t-tests with Bonferroni correction. However, those differences are not perceptible in practice.

Average performance scores on GR_{Ch} are similar to ML_{Ch} scores, but show higher variances. Notably, different RAs excel on GR_{Ch}, as reflected in a low Kendall's τ of 0.1282 between the rankings of RAs on ML_{Ch} and GR_{Ch}, according to nDCG@10. This highlights a distinct shift in algorithm effectiveness between datasets, suggesting that factors specific to GR_{Ch} impact what algorithms perform better in ways not observed in ML_{Ch}. Unlike ML_{Ch}, we note significant differences in performance between RAs on GR_{Ch}, with exceptions being non-significant differences in nDCG between UserkNN and ItemkNN, between PMF and MostPop in terms of MRR, and between UserkNN and ItemkNN, and MF and PMF in terms of MAP. Nevertheless, most differences are negligible in practice and the nDCG, MRR, and MAP scores remain overall low for the RAs.

On ML_{Ch} and GR_{Ch}, MostPop stands out as the lowest-performing, besides Random (according to paired t-tests with Bonferroni correction on all metrics). CB performs significantly worse than personalized RAs across all metrics (except for FunkSVD, MF, and AMF on MRR). CB and MostPop exhibit similar performance, making them the least effective when children are the main audience.

5 Discussion

Here, we reflect on the results reported in Section 4 and discuss their implications for the design, development, and evaluation of recommender systems.

5.1 Answer to RQs

Regarding **RQ1**, we observe that a high number of items in the item catalog include stereotypical representations. We find that Gender stereotypes are the most frequent, followed by Race stereotypes, with a relatively low occurrence of Religion stereotypes. This trend is consistent across all SDMs, which detect different aspects of stereotypes as described in Section 4.1. The predictions across different SDMs do not strongly overlap (Table 4a), indicating that different items are often identified as stereotypical based on different manifestations of stereotypes.

When examining how RAs reflect the frequency of stereotypes in their recommendations (**RQ2**), we observe that RAs commonly suggest stereotypical items to children (see Section 4.2). This tendency is not exclusive to any particular RA, but rather a pervasive trend across all. Using diverse metrics to gauge stereotype presence among suggestions, we compose an overview of the potential risks associated with this tendency for the vulnerable population studied: Not only is there a high frequency of stereotypical items in the recommendation lists but, more importantly, these items often appear high in the ranking. This is concerning because children typically engage more with the top-ranked items [27, 42, 71], translating to more frequent consumption of stereotypical content. Gender stereotypical items appear in almost every recommendation list, exacerbating concerns related to children's perceptions of Gender roles [23, 67, 105]. Although less prominent, Race and Religion stereotypes also commonly appear among recommendations, exposing children to unfavorable representations of those social groups.

The performance analysis indicates similar trends across different personalized RAs. We found no clear evidence for a relationship or trade-off between performance and stereotype presence in the recommendations (**RQ3**). Except for CB which performs the worst while being among the RAs that recommend the most stereotypes consistently, including the most Religion stereotypes, RAs generally exhibit similar behavior in terms of performance and stereotypical suggestions. On GR$_{Ch}$, we find that FunkSVD and PMF show tendencies towards higher presence of some stereotypes, and are among the worse-performing algorithms. However, these effects are unique to a single dataset and differences in performance are mostly negligible. It is questionable whether such differences would actually make a difference in terms of relevance for a user [32].

5.2 Implications

The widespread presence of stereotypes in recommendations across datasets, regardless of the detection method used, and the comparable performance of RAs, suggest that the presence of stereotypes in recommendations is not influenced by the specific methodology driving the recommendation generation process, such as neighborhood-based versus neural network-based approaches. One might assume that certain recommendation strategies would be more likely to pick up on specific properties of items that tend to be stereotypical, much like how certain algorithms are more prone to recommending popular items than others [2, 108]. We probed a wide range of algorithms, based on different methodologies to gain such insights. However, our results do not provide direct indications for such effects. Rather, RAs seem to consistently reflect the underlying stereotypes in the data sources utilized by RAs.

Throughout our analysis, we discovered that the issue of stereotypical recommendations is a complex topic, influenced by numerous underlying factors that contribute to the promotion of stereotypes. While performance may play a role in stereotype exposure, its extent remains uncertain, since there are no clear and consistent results across stereotypes and datasets. MostPop also tends to suggest stereotypical items frequently, suggesting a link between popularity and stereotype presence. This is alarming, given MostPop's frequent usage on online platforms, e.g., best-selling lists or top charts [17]. As many RAs are susceptible to popularity bias [2], MostPop's trend might be echoed in other RAs. Nonetheless, we also found inconsistencies for MostPop, as for some stereotypes, the inclusion of stereotypical items in its recommendations is significantly lower than for all other RAs. Overall, the high presence of stereotypes in the item catalogs shapes their presence in recommendations. Still, there are factors—beyond the selection of the algorithm—that could further impact stereotype inclusion in recommendations. Of note, there is a wide range of domains and media types available on online platforms, ranging from videos to news, which might impact the effects and differ strongly from the tendencies that we explored with the two datasets. Children might interact differently with platforms and items available, a facet that we are not able to explore in this fully algorithmic exploration. Clearly, no property of the recommendation setting in isolation can fully explain the stereotype promotion phenomenon. Rather, many aspects collectively influence the outcomes and the associated harm potentially experienced by children. Understanding these complex dynamics is vital to promoting ethical recommendations [99] and making informed decisions about how to address stereotypes in recommendation systems.

Stereotypes are a nuanced and multifaceted phenomenon; making automatic stereotype detection a complex task. In this work, we use text descriptions of items to infer whether the respective items are actually stereotypical. Informed by our examination of predictions based on movies and book descriptions, we posit that Gender stereotypes might not only be more prominent, but they might also be relatively easier to identify than Race and Religion; i.e., the gender of a character can often be assumed from pronouns for a simplified two-gender approach, whereas social groups are seldom explicitly stated in descriptions. We also noted that stereotypes require a broader context of the media content for accurate detection; for instance, in the case of the representation of the crows in *Dumbo* (Figure 1). Although the crows are never explicitly labeled as a social group, their depiction suggests an association with black stereotypes, a representation widely criticized [88]. Hence, SDMs must go beyond relying on simple labeling and textual data to consider more nuanced aspects. Mindful of this fact, we leveraged GPT to further probe textual data [46, 58] and potentially derive underlying, implicit stereotypes that an item might possess, but that are not explicitly mentioned in the item description. However, methods based on LLMs are not free from limitations. In our analysis, the main one is that LLMs operate as a black box, lacking direct insights into their reasoning. This makes it challenging to validate their reliability and capabilities, and their results are less interpretable. To get a more realistic picture of how stereotypes propagate through RAs, a combination of transparent methods—like NGIM or lexicon models [22] that can easily be explained and reasons for their classification can be followed—and more complex ones capable of identifying implicit manifestations of stereotypes is necessary. While no comprehensive method for universal stereotype detection currently exists, simultaneously considering various complementary methods and weighing their trade-offs is a step toward a more thorough understanding. Stereotypes manifest in different ways and have different implications that need to be evaluated carefully.

Looking ahead, progress in detection methods should involve refining both explicit and implicit identification techniques and enhancing stereotype recognition across a wider array of media and user contexts to increase the generalizability of our findings. Moreover, it is crucial to understand how stereotypes affect different user groups, particularly children, by considering their developmental stages, experiences, and social contexts. By integrating more refined insights about items and users into SDMs, we can better assess the specific harms that certain content might pose to children. This, in turn, lays the foundation for developing targeted harm mitigation strategies tailored to individual needs and vulnerabilities.

Reflections from our results highlight the critical importance of developing effective mitigation strategies and incorporating stereotype considerations into recommender system design. This is especially crucial for systems intended for broad audiences, as children are inevitably exposed to automated suggestions, amplifying the potential for harm caused by stereotypes. As stereotype detection advances, the potential to implement more refined, nuanced, and context-aware mitigation strategies will grow, helping to reduce the negative impacts on children and other vulnerable user groups. Initial mitigation approaches, such as re-ranking methods [e.g., 110] that deprioritize stereotypical content or content filters, may offer a starting point. Implementing initial methods, focusing on certain stereotypes instead of aiming to avoid every possible risk factor can already contribute to safeguarding children. In the future, such methods could become more and more complex and protective. Better understanding and the development of solutions regarding the harmful impacts of stereotypes and RAs is an iterative and cumulative process that benefits from continuous contributions and evolving perspectives. Popularity bias in recommender systems, for instance, was recognized as an issue as early as the 2000s. However, most research on the topic has emerged only in recent years and from various perspectives [53], illustrating how a single issue can be viewed and targeted in distinct ways. For example, studies have examined its effects on providers [25] and consumers [2, 108], and the research has evolved to adopt multi-stakeholder approaches [1]. Nowadays various popularity bias mitigation techniques exist that deal with popularity bias in different ways [53]. Similarly, addressing the harms posed by RAs through stereotype presentation to children demands a shift toward broader, multi-faceted insights and ongoing collaboration across fields.

5.3 Considerations for Child-Aware Recommender Systems Design

We consider this work a starting point for benevolent recommender system design, one that considers aspects that can impact a child's well-being. While there are various other potential harms and fine-grained properties of children and the recommendation setting that could impact children [102], we provide insights into one specific risk. Children's exposure to RAs' decisions on online platforms could have far-reaching implications in light of self-stereotyping. Children might see themselves reflected in these media representations—serving as a mirror through which they view the world and themselves. For example, gender-stereotypical content seldom associates girls with activities in the STEM field, negatively affecting their perception of their abilities and motivation to engage in this field, which perpetuates traditional gender roles [23, 67, 105]. Additionally, the gender imbalance [95, 96] highlighted by the frequent recommendation of items featuring predominantly male characters, can limit girls' ability to relate to characters, shaping their media experiences and perceptions [47]. Moreover, stereotypical representations can be mirrored by children and fuel prejudice and harmful attitudes toward groups; e.g., U.S. children often hold negative stereotypes about Arab Muslims [16], reflecting the portrayal of these groups in American media [94]. As children form attitudes based on media exposure, RAs can perpetuate such harmful stereotypes about various social groups, impacting children's perceptions, attitudes, beliefs, and behaviors [11]. This results in a systematic propagation of stereotypes on a societal level, impacting individuals from childhood into adulthood, and leading to reinforcement of societal stereotypes. Essentially, the influence of RAs extends far beyond mere suggestions, shaping not only individual perceptions but also societal norms, underscoring the urgent need for proactive observation, research, and interventions.

Putting our insights into practice presents a complex challenge. While the prevalence of stereotypical items in recommendations highlights the need to protect children, the actual harm posed by such items is highly context-dependent. The same item may have different implications depending on factors such as the target user, the platform it appears on, and the cultural context, highlighting the nuanced nature of stereotypical representations and, thus, the challenges in detecting and mitigating them. At the same time, not all stereotypes inherently lead to negative impacts. Self-stereotyping can foster a sense of belonging to a social group, contributing to the development of social identity, which can have positive effects, such as emotional attachment and even motivating disadvantaged groups to drive social change [98]. Furthermore, being aware of stereotypical representations can help reverse negative effects [61]. By empowering children to recognize and reflect on current stereotypes in media, media literacy can be fostered, which is crucial for addressing and challenging existing stereotypes. This awareness enables children to actively critique and intervene in the formation of stereotypes [90]. Take Disney's Dumbo as an example; instead of removing the movie from their platforms, Disney links information about the apparent stereotypes and encourages viewers to reflect on them[14]. Similarly, we believe that removing Example 2, presented in Section 4.1, would not be an appropriate strategy to safeguard children, despite it being deemed stereotypical by GPT. The stereotypes evident in this story could significantly contribute to children's understanding and reflection on the historical suppression of people of color. However, it has to be ensured that the child is able to comprehend such stories and reflect on them.

These examples underscore considerations that need to be made when designing recommender systems that balance protecting children from harmful stereotypes with fostering opportunities for critical reflection and learning. Practically, this means they should not adopt a blanket approach of removing all stereotypical content. Instead, they should be equipped to evaluate the context in which stereotypes appear and assess their potential impact on children. Recommender systems that are aware of children as their users have the power to foster media literacy and empower users to critically reflect on stereotypes. Advancements in detection methods that are tailored to different maturity levels result in recommender systems that can balance between removing unfitting

[14] https://storiesmatter.thewaltdisneycompany.com/

content and providing additional context and learning material, thus serving as a tool for promoting awareness and social growth while safeguarding children from undue harm.

Achieving practical progress in equipping recommender systems with these abilities cannot only be accomplished by the RecSys research community but requires a collaborative effort among stakeholders who can influence or enforce children's safety. Computer scientists may take the lead in designing and validating methods to detect stereotypes and mitigate stereotype impacts. However, their efforts must be continuously guided by ethicists and the social science community offering insights into the harms of stereotypes. Companies responsible for platforms used by children must implement measures that align with regulatory standards set by policymakers. Additionally, caretakers and parents play a crucial role by sharing their awareness of children's online habits and the harm they may encounter.

In practice, facilitating such multi-perspective considerations of harmful content is a complex and demanding endeavor. One important step toward fostering conversations and action by different stakeholders is to enable platforms where these topics can be discussed and collaborations can thrive. Workshops, special issues in academic journals, and interdisciplinary conferences that openly encourage controversial and challenging conversations can provide valuable forums for these dialogues. Recent platforms, such as the RecSoGood workshop [13] or the 'IR for Good' track at the 'European Conference on Information Retrieval' [68] highlight the importance of sustainable and social goals of recommender systems, and the IR4U2 workshop focuses on the experiences of understudied users specifically [81]. These platforms not only allow the exchange of insights and foster partnerships between people from different perspectives and backgrounds, but they also serve to draw attention to stakeholders who could make a significant impact yet may not be aware of the issues at stake. Creating these collaborative spaces paves the way for a shared understanding of the challenges involved, allowing for the development of actions that can be applied meaningfully in real-world settings.

The presentation of harmful content by RAs—-beyond stereotypical content—carries inherent risks to young audiences. Thus, the need for collaborative, actionable efforts to address this issue is pressing. Early attempts at direct and proactive solutions are beginning to emerge, such as the risk assessment proposed in the EU's AI Act, which requires platforms utilizing high-risk systems to evaluate potential harms before content reaches users [107]. While risk assessments represent an essential step forward, they remain in development and require a deeper understanding of the harms at hand. These initial regulatory efforts underscore the importance of a structured, preventive approach, yet their effectiveness will depend on continued evolution, practical implementation, and alignment with insights from multi-stakeholder discussions. By combining these structured assessments with collaborative input, we can move toward building recommender systems that prioritize safety and well-being, especially for younger users.

5.4 Limitations

Our experimental exploration framework produced results that allowed us to generate valuable insights, but it is not without limitations. We studied the problem on 2 datasets, one with only 218 child profiles (ML). This sample may not fully capture the diverse experiences of all children, but it provides valuable indications for apparent effects. In addition, GR does not include "real" user demographics. Instead, we enabled our exploration by making assumptions about the consumption behavior of users in order to gather children's profiles. Although our SDM approaches were constrained to certain social groups, including a two-gender approach and only limited religious groups, they still laid the groundwork for understanding the impacts of stereotypes conveyed by RAs to children. Together, they yielded a comprehensive overview of stereotypes in the data and recommendations. We recognize that item descriptions provide only one lens through which to model media content and that books and movies are not the sole types of items presented on online platforms accessed by children. We limited the scope of our exploration to gain initial insights and spark reflection on the topic of stereotype exposure to children. Future work should expand our proposed framework to incorporate aspects besides textual data like audio, images, and video as these could also influence children's consumption patterns.

As a purely empirical study, we do not directly capture children's actual perceptions. Instead, our analysis is scoped to gain insights into the ability of RAs to disseminate stereotypical content, motivated by the potential impact such content might have on children. Numerous studies have examined how stereotypical media influences perceptions [6, 21, 105], yet few consider the role of recommender systems in presenting such content. Observing the full interplay between stereotypical media, recommender systems, and children's perceptions would require careful ethical consideration and approval, which could provide deeper insights into this complex dynamic.

6 Conclusion

Recognizing the pivotal role recommender systems play in shaping users' experiences in online environments, we investigated the prevalence of stereotypical items in RAs' suggestions to children. Up to now, there is no indication of whether children's exposure to stereotypes through recommender systems is a challenge that needs to be considered in recommender systems research. Our research documents the omnipresence of stereotypes in media that children might receive and emphasizes the mirroring of such stereotypes in recommendations and, thus, the dissemination of potentially harmful content by state-of-the-art RAs employed in use cases where children can be affected.

We highlight that children can be exposed to stereotypes as a result of their direct or indirect interactions with RAs—whether they are the main users of a system or whether they are just passively exposed to its decisions. This fact influences how children will perceive the media that they consume, and how it, in turn, will impact them. Hence, we argue that such considerations should be a necessity when we evaluate whether a recommender system *serves* a child. Serving the needs and interests of a child is not just performance, as it is not for any user [32]. Particularly children are in a vulnerable and impressionable phase, where various aspects could have an immense impact—stereotypes being one of those.

Our work serves as a foundation for future research toward an even better understanding of the risks RA can pose to children and—ultimately—how to mitigate negative effects. Building on the findings from this work, we aim to spark discussions and encourage research that informs the design of novel RAs capable of protecting children from negative exposure to stereotypical content online. As argued by Bigler and Liben [10], the perpetuation of stereotypes is subject to social control, suggesting that the development of mitigation strategies and filters to avoid harmful stereotype exposure, as well as recommending materials that actively challenge stereotypes, could aid in diminishing their influence on children. This approach may contribute to the eventual elimination of stereotypical attitudes and behaviors [54]. By advancing knowledge regarding what is actually recommended to children, we contribute to laying a much-needed foundation for safeguarding children's well-being and actually *serving* their needs.

References

[1] Himan Abdollahpouri and Robin Burke. 2021. Multistakeholder recommender systems. In *Recommender systems handbook*. Springer, 647–677.

[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The unfairness of popularity bias in recommendation. *RMSE workshop, in conjunction with the 13th ACM Conference on Recommender Systems (RecSys 2019), Available at https://doi.org/10.48550/arXiv.1907.13286* (2019).

[3] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. 54–59.

[4] Vito Walter Anelli, Alejandro Bellogín, Tommaso Di Noia, Dietmar Jannach, and Claudio Pomo. 2022. Top-n recommendation algorithms: A quest for the state-of-the-art. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 121–131.

[5] Vito Walter Anelli, Alejandro Bellogín, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. 2021. Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 2405–2414.

[6] Markus Appel and Silvana Weber. 2021. Do mass mediated stereotypes harm members of negatively stereotyped groups? A meta-analytical review on media-generated stereotype threat and stereotype lift. *Communication Research* 48, 2 (2021), 151–179.

[7] Albert Bandura. 2009. Social cognitive theory of mass communication. In *Media effects*. Routledge, 110–140.

[8] Albert Bandura and Richard H Walters. 1977. *Social learning theory*. Vol. 1. Englewood cliffs Prentice Hall.

[9] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems* 24 (2011).

[10] Rebecca S Bigler and Lynn S Liben. 2006. A developmental intergroup theory of social stereotypes and prejudice. *Advances in child development and behavior* 34 (2006), 39–89.

[11] Rebecca S Bigler and Lynn S Liben. 2007. Developmental intergroup theory: Explaining and reducing children's social stereotyping and prejudice. *Current directions in psychological science* 16, 3 (2007), 162–166.

[12] Dania Bilal. 2010. The mediated information needs of children on the Autism Spectrum Disorder (ASD). In *Proceedings of the 31st ACM SIGIR Workshop on Accessible Search Systems, Geneva, Switzerland*. ACM Geneva, 42–49.

[13] Ludovico Boratto, Allegra De Filippo, Elisabeth Lex, and Francesco Ricci. 2024. First International Workshop on Recommender Systems for Sustainability and Social Good (RecSoGood 2024). In *Proceedings of the 18th ACM Conference on Recommender Systems*. 1239–1241.

[14] Pedro Bordalo, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. Stereotypes. *The Quarterly Journal of Economics* 131, 4 (2016), 1753–1794.

[15] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 1664–1674.

[16] Christia Spears Brown, Hadeel Ali, Ellen A Stone, and Jennifer A Jewell. 2017. US children's stereotypes and prejudicial attitudes toward Arab Muslims. *Analyses of Social Issues and Public Policy* 17, 1 (2017), 60–83.

[17] Rocío Cañamares and Pablo Castells. 2018. Should I follow the crowd? A probabilistic analysis of the effectiveness of popularity in recommender systems. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 415–424.

[18] Tessa ES Charlesworth, Victor Yang, Thomas C Mann, Benedek Kurdi, and Mahzarin R Banaji. 2021. Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science* 32, 2 (2021), 218–240.

[19] Linda Charmaraman, Alicia Doyle Lynch, Amanda M Richer, and Jennifer M Grossman. 2022. Associations of early social media initiation on digital behaviors and the moderating role of limiting use. *Computers in Human Behavior* 127 (2022), 107053.

[20] Cinemagoer. 2024. *Cinemagoer*. https://cinemagoer.github.io/

[21] Sarah M Coyne, Jennifer Ruh Linder, Eric E Rasmussen, David A Nelson, and Kevin M Collier. 2014. It'sa bird! It'sa plane! It'sa gender stereotype!: Longitudinal associations between superhero viewing and gender stereotyped play. *Sex Roles* 70 (2014), 416–430.

[22] Jenna Cryan, Shiliang Tang, Xinyi Zhang, Miriam Metzger, Haitao Zheng, and Ben Y Zhao. 2020. Detecting gender stereotypes: Lexicon vs. supervised learning methods. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–11.

[23] Dario Cvencek, Andrew N Meltzoff, and Anthony G Greenwald. 2011. Math–gender stereotypes in elementary school children. *Child development* 82, 3 (2011), 766–779.

[24] Tommaso Di Noia, Roberto Mirizzi, Vito Claudio Ostuni, Davide Romito, and Markus Zanker. 2012. Linked open data to support content-based recommender systems. In *Proceedings of the 8th international conference on semantic systems*. 1–8.

[25] Karlijn Dinnissen and Christine Bauer. 2023. Amplifying artists' voices: Item provider perspectives on influence and fairness of music streaming platforms. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. 238–249.

[26] Nicola Döring and M Rohangis Mohseni. 2019. Fail videos and related video comments on YouTube: a case of sexualization of women and gendered hate speech? *Communication research reports* 36, 3 (2019), 254–264.

[27] Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Serdyukov. 2010. Query log analysis in the context of information retrieval for children. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 847–848.

[28] Michael Ekstrand. 2017. Challenges in evaluating recommendations for children. In *International Workshop on Children & Recommender Systems. Available at: shorturl. at/osFV9*.

[29] Abir El Shaban. 2017. Gender stereotypes in fantasy fairy tales: Cinderella. *AWEJ for Translation & Literary Studies, Volume* 1 (2017).

[30] Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. 2020. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing & Management* 57, 6 (2020), 102377.

[31] Guglielmo Faggioli, Laura Dietz, Charles LA Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*. 39–50.

[32] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM conference on recommender systems*. 101–109.

[33] Antske Fokkens, Nel Ruigrok, Camiel Beukeboom, Gagestein Sarah, and Wouter Van Atteveldt. 2018. Studying muslim stereotyping through microportrait extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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

[35] Simon Funk. 2006. Netflix update: Try this at home.

[36] Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem. 2022. Masked Language Models as Stereotype Detectors?. In *EDBT 2022*.

[37] David Garcia, Ingmar Weber, and Venkata Garimella. 2014. Gender asymmetries in reality and fiction: The bechdel test of social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8. 131–140.

[38] Emilia Gómez Gutiérrez, Vicky Charisi, and Stephane Chaudron. 2021. Evaluating recommender systems with and for children: towards a multi-perspective framework. In *CEUR Workshop Proceedings. 2021; 2955*. CEUR Workshop Proceedings.

[39] Goodreads, Inc. 2024. *Goodreads*. https://www.goodreads.com/

[40] Ulrike Gretzel and Daniel R Fesenmaier. 2006. Persuasion in recommender systems. *International Journal of Electronic Commerce* 11, 2 (2006), 81–100.

[41] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247* (2017).

[42] Jacek Gwizdka and Dania Bilal. 2017. Analysis of children's queries and click behavior on ranked results and their thought processes in google search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*. 377–380.

[43] David L Hamilton and Tina K Trolier. 1986. Stereotypes and stereotyping: An overview of the cognitive approach. (1986).

[44] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.

[45] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[46] Brittany Ho, Khanh Linh Nguyen, Manohar Dhulipala, Vivek Krishnamani Pallipuram, et al. 2024. ChatReview: A ChatGPT-enabled natural language processing framework to study domain-specific user reviews. *Machine Learning with Applications* 15 (2024), 100522.

[47] Cynthia Hoffner and Martha Buchanan. 2005. Young adults' wishful identification with television characters: The role of perceived similarity and character attributes. *Media psychology* 7, 4 (2005), 325–351.

[48] Tomasz Huk et al. 2016. Use of Facebook by children aged 10-12. Presence in social media despite the prohibition. *The New Educational Review* 46, 1 (2016), 17–28.

[49] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring misinformation in video search platforms: An audit study on YouTube. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–27.

[50] Israel Saeta Pérez. 2024. *Gender Guesser*. https://pypi.org/project/gender-guesser/

[51] Kenneth Joseph, Wei Wei, and Kathleen M Carley. 2017. Girls rule, boys drool: Extracting semantic and affective stereotypes from Twitter. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1362–1374.

[52] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*. 3819–3828.

[53] Anastasiia Klimashevskaia, Dietmar Jannach, Mehdi Elahi, and Christoph Trattner. 2024. A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction* (2024), 1–58.

[54] Ellen E Kneeskern and Patricia A Reeder. 2022. Examining the impact of fiction literature on children's gender stereotypes. *Current psychology* 41, 3 (2022), 1472–1485.

[55] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[56] Yehuda Koren, Steffen Rendle, and Robert Bell. 2021. Advances in collaborative filtering. *Recommender systems handbook* (2021), 91–142.

[57] Natalia Kucirkova. 2019. The Learning Value of Personalization in Children's Reading Recommendation Systems: What Can We Learn From Constructionism? *International Journal of Mobile and Blended Learning (IJMBL)* 11, 4 (2019), 80–95.

[58] Tiziano Labruna, Sofia Brenna, Andrea Zaninello, and Bernardo Magnini. 2023. Unraveling chatgpt: A critical analysis of ai-generated goal-oriented dialogues and annotations. In *International Conference of the Italian Association for Artificial Intelligence*. Springer, 151–171.

[59] Monica Landoni, Theo Huibers, Emiliana Murgia, and Maria Soledad Pera. 2024. Good for Children, Good for All?. In *European Conference on Information Retrieval*. Springer.

[60] Monica Landoni, Davide Matteri, Emiliana Murgia, Theo Huibers, and Maria Soledad Pera. 2019. Sonny, Cerca! evaluating the impact of using a vocal assistant to search at school. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*. Springer, 101–113.

[61] Lorella Lepore and Rupert Brown. 2002. The role of awareness: Divergent automatic stereotype activation and implicit judgment correction. *Social Cognition* 20, 4 (2002), 321–351.

[62] Elizabeth Levin. 2011. *Child Development*. Springer US, Boston, MA, 337–339. doi:10.1007/978-0-387-79061-9_523

[63] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*. 689–698.

[64] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing* 7, 1 (2003), 76–80.

[65] Sonia Livingstone, Kjartan Ólafsson, and Elisabeth Staksrud. 2011. Social networking, age and privacy. (2011).

[66] George F Luger. 2005. *Artificial intelligence: structures and strategies for complex problem solving*. Pearson education.

[67] Allison Master. 2021. Gender stereotypes influence children's STEM motivation. *Child Development Perspectives* 15, 3 (2021), 203–210.

[68] Graham McDonald, Craig Macdonald, and Iadh Ounis. 2024. Report on the 46th European Conference on Information Retrieval (ECIR 2024). In *ACM SIGIR Forum*, Vol. 58. ACM New York, NY, USA, 1–18.

[69] Ashlee Milton, Michael Green, Adam Keener, Joshua Ames, Michael D Ekstrand, and Maria Soledad Pera. 2019. StoryTime: eliciting preferences from children for book recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 544–545.

[70] Ashlee Milton, Emiliana Murgia, Monica Landoni, Theo Huibers, and Maria Soledad Pera. 2019. Here, there, and everywhere: Building a scaffolding for children's learning through recommendations. (2019).

[71] Ashlee Milton and Maria Soledad Pera. 2020. Evaluating information retrieval systems for kids. *4th International and Interdisciplinary Perspectives on Children & Recommender and Information Retrieval Systems (KidRec '20), in conjunction with ACM IDC 2020, Available at https://doi.org/10.48550/arXiv.2005.12992* (2020).

[72] Andriy Mnih and Russ R Salakhutdinov. 2007. Probabilistic matrix factorization. *Advances in neural information processing systems* 20 (2007).

[73] Emiliana Murgia, Monica Landoni, Theo Huibers, Jerry Alan Fails, and Maria Soledad Pera. 2019. The seven layers of complexity of recommender systems for children in educational contexts. (2019).

[74] 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)*. 5356–5371.

[75] 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)*. 1953–1967.

[76] Xia Ning and George Karypis. 2011. Slim: Sparse linear methods for top-n recommender systems. In *2011 IEEE 11th international conference on data mining*. IEEE, 497–506.

[77] UK OfCom. 2022. Children and parents: Media use and attitudes report 2022. *London: Office of Communications London* (2022).

[78] UK OfCom. 2023. Children and parents: Media use and attitudes report 2023. *London: Office of Communications London* (2023).

[79] Jahna Otterbacher, Jo Bates, and Paul Clough. 2017. Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proceedings of the 2017 chi conference on human factors in computing systems*. 6620–6631.

[80] Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. 2020. Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 522–533.

[81] Maria Soledad Pera, Emiliana Murgia, Monica Landoni, Federica Cena, Noemi Mauro, and Theo Huibers. 2024. Report on the 1st Workshop on Information Retrieval for Understudied Users (IR4U2 2024) at ECIR 2024. In *ACM SIGIR Forum*, Vol. 58. ACM New York, NY, USA, 1–5.

[82] Amifa Raj and Michael D Ekstrand. 2022. Fire Dragon and Unicorn Princess; Gender Stereotypes and Children's Products in Search Engine Responses. *Workshop on eCommerce (SIGIR eCOM'22), in conjunction with ACM SIGIR 2022, Available at https://doi.org/10.48550/arXiv.2206.13747* (2022).

[83] Amifa Raj, Ashlee Milton, and Michael D Ekstrand. 2021. Pink for Princesses, Blue for Superheroes: The Need to Examine Gender Stereotypes in Kid's Products in Search and Recommendations. *5th International and Interdisciplinary Perspectives on Children & Recommender and Information Retrieval Systems (KidRec '21), in conjunction with ACM IDC 2021, Available at https://doi.org/10.48550/arXiv.2105.09296* (2021).

[84] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI2009)*.

[85] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2021. Recommender systems: Techniques, applications, and challenges. *Recommender Systems Handbook* (2021), 1–35.

[86] Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. 74–79.

[87] Alan Said and Alejandro Bellogín. 2014. Comparative recommender system evaluation: benchmarking recommendation frameworks. In *Proceedings of the 8th ACM Conference on Recommender systems*. 129–136.

[88] Nicholas Sammond. 2011. Dumbo, Disney, and Difference: Walt Disney Productions and Film as Children's Literature. *The Oxford Handbook of Children's Literature* (2011), 147–66.

[89] 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*.

[90] Erica Scharrer and Srividya Ramasubramanian. 2015. Intervening in the media's influence on stereotypes of race and ethnicity: The role of media literacy education. *Journal of Social Issues* 71, 1 (2015), 171–185.

[91] Markus Schedl and Christine Bauer. 2017. Online music listening culture of kids and adolescents: Listening analysis and music recommendation tailored to the young. *1st International Workshop on Children and Recommender Systems, in conjunction with 11th ACM Conference on Recommender Systems (RecSys 2017), Available at: https://doi.org/10.48550/arXiv.1912.11564* (2017).

[92] Nick Seaver. 2019. Captivating algorithms: Recommender systems as traps. *Journal of material culture* 24, 4 (2019), 421–436.

[93] Avi Segal, Ziv Katzir, Kobi Gal, Guy Shani, and Bracha Shapira. 2014. Edurank: A collaborative filtering approach to personalization in e-learning. In *Educational Data Mining 2014*. Citeseer.

[94] Jack G Shaheen. 2003. Reel bad Arabs: How Hollywood vilifies a people. *The ANNALS of the American Academy of Political and Social science* 588, 1 (2003), 171–193.

[95] Stacy L Smith, Marc Choueiti, Katherine Pieper, Traci Gillig, Carmen Lee, and Dylan DeLuca. 2015. Inequality in 700 popular films: Examining portrayals of gender, race, & LGBT status from 2007 to 2014. *Media, Diversity, & Social Change Initiative* (2015), 1–29.

[96] Stacy L Smith, Katherine Pieper, and Marc Choueiti. 2015. Gender bias without borders: An investigation of female characters in popular films across 11 countries. (2015).

[97] Lawrence Spear, Ashlee Milton, Garrett Allen, Amifa Raj, Michael Green, Michael D Ekstrand, and Maria Soledad Pera. 2021. Baby shark to barracuda: Analyzing children's music listening behavior. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 639–644.

[98] Russell Spears. 2011. Group identities: The social identity perspective. In *Handbook of identity theory and research*. Springer, 201–224.

[99] Tiffany Ya Tang and Pinata Winoto. 2016. I should not recommend it to you even if you will like it: the ethics of recommender systems. *New Review of Hypermedia and Multimedia* 22, 1-2 (2016), 111–138.

[100] TMDB, Inc. 2024. *The Movie DB*. https://www.themoviedb.org/

[101] Ya-Lun Tsao. 2008. Gender issues in young children's literature. *Reading improvement* 45, 3 (2008), 108–114.

[102] Robin Ungruh and Maria Soledad Pera. 2024. Ah, that's the great puzzle: On the Quest of a Holistic Understanding of the Harms of Recommender Systems on Children. *DCDW, in conjunction with ACM IDC 2024, Available at: https://doi.org/10.48550/arXiv.2405.02050* (2024).

[103] María Cora Urdaneta-Ponte, Amaia Mendez-Zorrilla, and Ibon Oleagordia-Ruiz. 2021. Recommendation systems for education: Systematic review. *Electronics* 10, 14 (2021), 1611.

[104] Mengting Wan and Julian McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM conference on recommender systems*. 86–94.

[105] L Monique Ward and Petal Grower. 2020. Media and the development of gender role stereotypes. *Annual Review of Developmental Psychology* 2, 1 (2020), 177–199.

[106] Niobe Way, María G Hernández, Leoandra Onnie Rogers, and Diane L Hughes. 2013. "I'm not going to become no rapper" stereotypes as a context of ethnic and racial identity development. *Journal of Adolescent Research* 28, 4 (2013), 407–430.

[107] Steve Wood. 2024. Children and Social Media Recommender Systems: How Can Risks and Harms be Effectively Assessed in a Regulatory Context? *Available at SSRN 4978809* (2024).

[108] Emre Yalcin and Alper Bilge. 2022. Evaluating unfairness of popularity bias in recommender systems: A comprehensive user-centric analysis. *Information Processing & Management* 59, 6 (2022), 103100.

[109] Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420* (2023).

[110] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1569–1578.

Received 17 September 2024; revised 4 December 2024; accepted 26 February 2025