Tidying Up the Conversational Recommender Systems' Biases

Armin Moradi[1,2], Golnoosh Farnadi[1,2,3]

[1]Mila, Quebec AI Insitute
[2]Université de Montréal
[3]McGill University

Abstract

The growing popularity of language models has sparked interest in conversational recommender systems (CRS) within both industry and research circles. However, concerns regarding biases in these systems have emerged. While individual components of CRS have been subject to bias studies, a literature gap remains in understanding specific biases unique to CRS and how these biases may be amplified or reduced when integrated into complex CRS models. In this paper, we provide a concise review of biases in CRS by surveying recent literature. We examine the presence of biases throughout the system's pipeline and consider the challenges that arise from combining multiple models. Our study investigates biases in classic recommender systems and their relevance to CRS. Moreover, we address specific biases in CRS, considering variations with and without natural language understanding capabilities, along with biases related to dialogue systems and language models. Through our findings, we highlight the necessity of adopting a holistic perspective when dealing with biases in complex CRS models.

Keywords

Conversational Recommender Systems, Bias, Responsible AI, Large Language Models

1. Introduction

In recent years, conversational recommender systems (CRSs) [1, 2, 3, 4, 5, 6, 7, 8, 9] have garnered significant attention, reshaping personalized recommendations through interactive user engagements. This transformation is notably supported by the successful integration of large language models (LLMs) like ChatGPT [10, 11], thereby driving the widespread deployment of LLMs in various applications. Such models have found substantial integration in prominent platforms, including Microsoft Bing[1], which has invigorated dialogue search engines and recommender systems with unprecedented capabilities and paving the way for a new era of user engagement.

Although biases within recommender systems have garnered significant attention [12, 13, 14, 15, 16, 17, 18], the examination of biases in conversational recommender systems remains a relatively unexplored domain [19]. Despite the conceptual alignment between conventional recommender systems and their conversational counterparts, the latter exhibit heightened complexity and an increased potential for biases [20]. It is worth noting that some existing research has specifically addressed biases in conversational recommender systems [3, 4, 5, 6, 16, 20, 21]. However, no concerted effort has been undertaken to systematically categorize and analyze biases unique to conversational recommender systems, including how previously studied biases manifest within the conversational context. This paper aims to bridge this gap by offering a preliminary exploration into the intricate biases that characterize these complex systems.

This study conducts a systematic literature review to explore biases within conversational recommender systems. The methodology involves analyzing recent papers from top conferences in machine learning and information retrieval. Keyword searches within titles and abstracts identify relevant contributions that shed light on biases, including fairness concerns and bias amplification.

The paper begins by thoroughly investigating biases in classic recommender systems, establishing a foundation for understanding common biases, addressing similar notions to each of them, and most importantly, examining each bias in a conversational setting in Section 5. This step ensures a holistic grasp of biases across traditional recommendation systems. We then delve into each bias within CRSs. We start with focusing on CRSs without natural language understanding which uses basic dialogue systems for user interaction in Section 6. Furthermore, to capture diverse aspects and potential biases arising from natural language understanding, a dedicated literature review is conducted on the more complex CRSs that aim to understand natural language in Section 7. Following that, in Section 8, we present the limitations and potential avenues for future research of our study. Finally, in Section 9, we draw conclusions and wrap up the paper.

Fifth Knowledge-aware and Conversational Recommender Systems (KaRS) Workshop @ RecSys 2023, September 18–22 2023, Singapore.
armin.moradi@mila.quebec (A. Moradi); farnadig@mila.quebec (G. Farnadi)

[1]https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI's-GPT-4

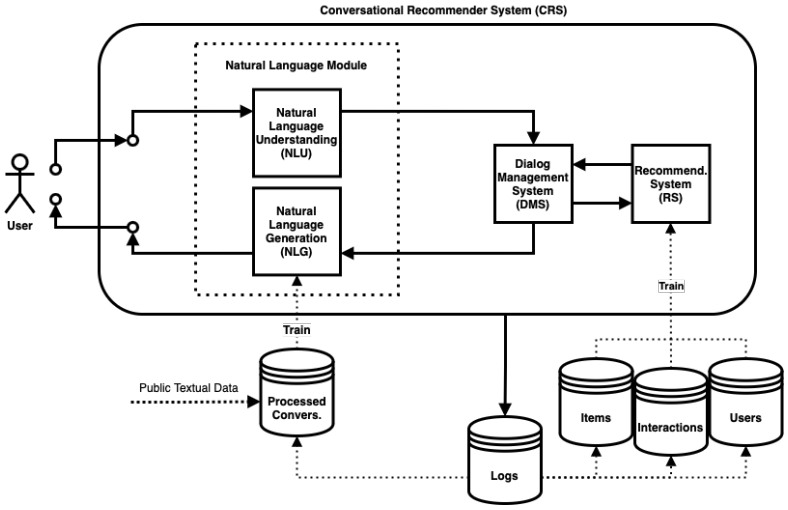

Figure 1: Architecture of the Conversational Recommender System. The diagram illustrates the intricate network of components, including a natural language module, dialogue management system, and recommender system working collaboratively to deliver personalized recommendations in a dynamic conversational interface.

2. Related Work

Although biases in general machine learning models, natural language processing (NLP), and recommender systems (RS) domains have undergone extensive study, biases in conversational recommender systems (CRSs) have received limited attention.

On a broad aspect, ML biases have been a subject of interest, with an example of the work by Mehrabi et al. [22] which provides a comprehensive survey on the challenges and approaches to mitigate biases in ML models. Additionally, NLP biases have received attention, and for instance, Blodgett et al. [23] discussed the manifestation and implications of biases in language models which can be seen as an important part of CRSs.

Moreover, several surveys have focused on biases and fairness in recommender systems (RS) as a whole [12, 24, 25, 13]. These surveys provide valuable insights into biases present in RS, which is a sub-module of CRSs, therefore they can be used as valuable sources for specific investigation in a conversational setting.

Within the realm of CRS biases, recent research has shed light on different biases that can arise in such systems. For example, unintended biases are discussed by Shen et al. [21]. Lin et al. [26] quantified biases in CRS by exploring the fairness of recommendations across different user groups. Another study by Lin et al. [20] highlighted the importance of addressing biases in CRS to ensure equitable and inclusive recommendations.

As for CRS surveys, Jannach et al. [19, 27, 9] have contributed significantly to the understanding of conversational recommender systems. However, to the best of our knowledge, a dedicated survey paper focusing solely on the biases in conversational recommender systems is still missing from the literature. Consequently, this paper aims to embark on the initial journey towards gaining a deeper understanding of the biases present in CRS and their interactions, elucidating how biases from various components within this complex system can either be accentuated or alleviated.

3. Methodology

The primary objective of this study is to conduct a review of existing literature on potential biases in conversational recommender systems. To achieve this goal, a systematic approach is adopted, focusing on papers published in prominent machine learning and information retrieval-related conferences from January 2019 to June 2023. These conferences include knowledge discovery and data mining (KDD), Special Interest Group on Information Retrieval (SIGIR), ACM Conference on Recommender Systems, User Modelling, Adaptation and Personalization (UMAP), International World Wide Web Conference (WWW), Neural Information Processing Systems (NeurIPS), International Conference on Machine Learning (ICML).

To identify relevant papers, we conducted keyword searches within the titles or abstracts of the papers presented at these conferences. The selected keywords include bias, dialog, conversation, chat, question, mitigat*, recommend*, amplif*, fair*. These were chosen to align with key aspects within the scope of biases in conversa-

tional recommender systems and related notions such as dialogue systems and classic recommender systems. These keywords were thoughtfully chosen to reflect the themes within the scope of biases in conversational recommender systems. Specifically, they align with key aspects such as recommender systems, dialogue systems, fairness considerations, and the amplification of biases. The chosen keywords collectively encompass a wide spectrum of research that pertains to these facets, ensuring that our selection is representative of the relevant literature landscape. Subsequently, we filtered the results once again based on their contributions to the trustworthy intricacies of recommender systems, conversational recommender systems, or dialogue systems.

4. Conversational Recommender Systems

To explore biases within conversational recommender systems (CRSs), a precise definition and understanding of their model architecture are crucial. A Conversational Recommender System (CRS) is intricate, with interconnected components (Figure 1). Users interact with the system to either provide answers (feedback) to the system's questions (recommendations) or receive personalized recommendations. Within the CRS, a language module assumes a pivotal role, consisting of two submodules: natural language understanding (NLU) and natural language generation (NLG). The NLU empowers the system to understand user intentions, extracting insights from their input and prior profiling data. On the other hand, the NLG crafts coherent and contextually relevant responses in natural language. Alongside this, a recommender system undertakes user queries, harnessing available data to generate personalized recommendations.

The dialogue management system (DMS) as the heart of the model, is in charge of orchestrating conversations between the user and the system, ensuring logical flow and pertinence in each interaction. At each point of the dialogue, The DMS decides when to finalize recommendations or seek more information, guided by NLU and the recommender system and by navigating the outputs through the NLG. Continual system enhancement is achieved through the accumulation of conversation logs into a database, leveraging these processed logs to train both the natural language module and the recommender system. This iterative training process augments their capabilities over time, refining user interactions.

It's important to emphasize that the architecture outlined above represents the general structure of a CRS. However, certain CRSs do not incorporate natural language understanding and generation into their operations; instead, they adopt simplified input processing methods. This leads to the categorization of CRSs into two types: *Topic-guided CRSs*, which utilize natural language understanding, and *Attribute-aware CRSs*, which rely on simplified input processing methods and attribute inquiries, as classified by Ren et al. [5].

Another significant point to consider is that a traditional recommender system can also be illustrated by utilizing a subset of the modules featured in Figure 1. By removing the Natural Language Module and the Dialogue Management System (DMS) components, and by establishing a direct communication path connecting the Recommender System and the User, we can achieve a simplified structure. This approach facilitates a clearer understanding of the underlying nature and impact of biases discussed in Section 5.

5. Biases in Classic Recommender Systems

In this section, we examine biases in classic recommender systems across three key aspects for each bias. For each bias, first, we define and discuss the bias, drawing from relevant literature for a strong foundation. Second, under *Similar Notions*, we identify related notions that fall within each bias. And third, in *Through the CRS lens*, we analyze each bias in the context of conversational recommender systems, exploring specific works in this domain. This approach offers a wide-ranging perspective on biases in classic recommender systems within conversational interactions.

5.1. Popularity Bias

Popularity bias in recommender systems prioritizes popular items, assuming they are more likely to interest users. However, this bias can lead to a lack of diversity, overshadowing lesser-known options. Recognizing and mitigating popularity bias is crucial for developing recommender systems that offer a wider range of choices and promote serendipitous discovery [25, 28]. Besides investigating and mitigating this bias in a classic fashion, there are some other works that focus on different settings and notions. For example, Zhu et al. [17, 29] try to solve the popularity bias in a dynamic environment setting. Abdollahpouri et al. [30] see how different types of users are affected by popularity bias and in the study by Zhu et al. [17] the focus pertains to the challenge of ranking a set of equally favored items based on their popularity.

Similar notions:

- **Long-tail bias:** The opposite of popularity bias can also be the issue in a model, over-recommending the niche items instead of already established items [31].

- **Filter bubbles, Echo chambers, Polarization:** These concepts constitute a significant area of research within recommender systems and often arise as consequences of existing popularity bias. In the work by Michiels et al. [32], the filter bubble is characterized by a "decrease in any dimension of diversity." It's essential to recognize that this notion doesn't solely originate from popularity bias; various other biases can contribute as well. Wang et al. [33] present a user-controllable model to alleviate filter bubbles, which holds potential for application in conversational settings.

Through the CRS lens: Lin et al. [20] examine the prevalence of popularity bias by investigating three CRS models as baselines, revealing its existence in these systems as well. It is speculated that this bias can be alleviated through the conversational setting because the user is capable of criticizing the recommendations in order to have a more niche recommended list of items.

5.2. User Log Bias

Different underlying biases can lead to a biased log of user data. Therefore we define User Log bias as an umbrella term for the discrepancy between real-world representation of user-item interactions (the ground truth) and the recorded data. This variation can happen due to many reasons, such as oversimplifying the user logs, not having access to certain user data and imprecise user-item interaction modeling.

There are some works that address this bias with different approaches and different but similar bias definitions. Frumermann et al. [34] investigate the real meaning behind rejected items and whether all the rejected items should be seen as the same. On the same issue, Nazari et al. [35] make an effort to use user-implicit signals and Xu et al. [36] try to leverage unclicked items in the dataset in addition to the interactions. Lastly, Zhang et al. [37] focus on how we should tackle user inattentiveness to the items that are being interacted with but do not necessarily correlate with user satisfaction.
Similar Notions:

- **Conformity Bias:** It is a cognitive bias which is defined as users' disinclination towards negatively rating an item because of the item's high ranking or popularity [25, 38].
- **Exposure Bias:** Since each user is randomly exposed to a subset of items during her lifetime, her profiling is biased towards the items that they have already been exposed to [14, 39, 40]. Also, it indicates that unobserved items do not always represent negative preference [25].
- **Selection Bias:** It is a similar notion to exposure bias, as it is defined as the un-randomness of the missing interactions [41, 42, 15].
- **Exploitation Bias:** It is investigated that users have a tendency to interact with or rate items that they personally prefer [43, 44].

Through the CRS lens: Likewise, biases in conversational settings concerning user logs have been examined, particularly regarding the process of dataset curation. Szpektor et al. and Yu et al. [45, 46] note that the limited number of individuals responsible for labeling conversation quality may introduce bias stemming from their personal preferences. Additionally, Pang et al. [47] highlight their focus on cultural differences, which can result in divergent labelings of CRS datasets.

Looking back to the *similar notions*, conformity bias is also studied in the conversational recommender systems [48, 49]. Overall, the conversational nature of CRSs and the inevitable increasing complexity can cause the model to have more biases in the log than the classic models leading to a reduction of the quality of the datasets which needs to be addressed in future studies.

5.3. Recommendation Evaluation Bias

In order to evaluate a recommender system, certain priorities need to be addressed with respect to the needed specifications of the system in addition to common accuracy and ranking evaluation metrics. For example, serendipity and diversity of the recommendations and long-term vs short-term fairness can all be taken into account as a way to measure the recommender system [50, 51, 52]. Therefore, it is important to establish metrics to assess the various aspects or qualities of a recommender system that need to be evaluated. Also, the flaws of some of the already established metrics have been challenged [53, 54] and some even propose a metric-less offline evaluation method [55]. Additional metrics such as evaluating fairness in ranking [56], and recommendation uncertainty [57] are also addressed in the literature. Building on fairness, De et al. [58] go beyond fairness metrics and suggest that search engines can manipulate users while maintaining top-notch fairness metrics. Lastly, Wang et al. and Zhang et al. [59, 37] try to investigate user attention and how it affects recommendation models and the way users interact with it, directly influencing the ways we measure their capability.
Through the CRS lens: There can be various use-cases for conversational recommender systems and different measurements can and should be prioritized depending on them. This makes choosing the evaluation metrics for conversational recommender systems a challenging task. For example, Lin et al. [20] propose that the Success Rate metric does not indicate how much the recommender system is benefiting each of its individual users. When evaluating a topic-guided conversational recommender

system, specific metrics could come into play, such as psychologically inspired measures and those assessing conversation quality and engagement [60]. These metrics also address sub-objectives like accurate user satisfaction estimation [2, 61]. Lastly, it is very important to evaluate the natural-language-understanding CRSs in order to measure the gap between the understanding of the NLU and how much this understanding is utilized by the recommender systems [62]. One of the fundamental reasons is explored in the study by Zho et al. [8], where they investigate the disparity between the natural language representation of a potentially recommended item and its lack of precise alignment.

In the realm of CRS evaluation, the incorporation of new aspects amplifies complexity. Evaluating different modules and the complete CRS requires meticulous consideration, making CRSs more susceptible to recommendation evaluation biases.

5.4. Attribute Bias

Attribute bias is a concerning issue in recommender systems. These systems can inadvertently amplify existing societal biases by making recommendations based on certain sensitive attributes that each user can have, such as gender, race, or age. This bias can lead to unfair and discriminatory outcomes, as individuals from certain demographics may be systematically excluded or receive less favourable recommendations. Addressing demographic bias in recommender systems is essential for ensuring equal and equitable treatment for all users, regardless of their demographic characteristics.
Similar notions:

- **Demographic bias** and **minority bias** can also be utilized to refer to Attribute bias [63, 64].
- **User Activity Bias:** There are several papers discussing the disparate impact of active users on the model [65], and how differently the model interacts with them [66].
- **Sentiment Bias:** It is investigated that the more the users have positive interactions with the system they are more likely to get higher quality recommendations [16].

Through the CRS lens: Due to the potential usage of natural language processing in conversational models, other attributes of users can be exposed to the model and can be exploited. For instance, in a voice dialogue system, the accent of the user is a sensitive attribute that should ideally not influence the system's decision-making [67]. Also in the work by Cogswell et al. [68], it is discussed that the modality of presenting the data can affect the minorities' ability to perceive it. On a similar issue, different people interact differently with regard to their age and how experienced they are in interacting with a recommender system [69] and the model should be robust in different interactions and be able to extract the users' needs without emphasizing their demographic attributes.

In a conversational setting, user interaction empowers them to navigate and refine recommendations, allowing them to mitigate attribute biases in the final list, similar to how they can address popularity bias. Nevertheless, additional research is essential to examine which users' attribute biases can be effectively mitigated through conversation, and also to understand which biases, such as gender or race bias, might be exacerbated due to the inherent biases of various components within CRS.

5.5. Position Bias

It happens as users tend to interact with items in higher positions of the recommendation list regardless of the items' actual relevance so that the interacted items might not be highly relevant [25, 70].
Similar Notions:

- It is in accordance with **lead bias** in a work by Zhu et al. [71] for news recommender systems that show how the 'lead' part of the news in the recommender systems can bias a user's behavior towards the item.

Through the CRS lens: In a conversational setting, there are options to counter position bias, such as providing explainability for each recommended item and framing the recommendations using natural language. These strategies can potentially mitigate the impact of position bias. It's important to note that this bias is intertwined with Framing bias (Section 7), as the ranking can be seen as a framing of data presentation and position bias can be considered a form of framing bias as well. Nevertheless, it is crucial to acknowledge that explanation methods which rely on an additional or surrogate model to provide justifications for why specific items are recommended to the user and ranked higher, are also susceptible to biases. Consequently, these explanations might not accurately reflect the performance of the original model.

5.6. Personalization Bias

Personalization bias in recommender systems refers to the tendency of these systems to continuously recommend similar content based on a user's preferences, potentially limiting their exposure to diverse perspectives and new experiences. Balancing personalization with serendipity is crucial to mitigate this bias and ensure users are presented with a broader range of recommendations [25].
Similar Notions:

- **User preference Amplification:** Kalimeris et al. [72] talk about how even relevant and high-quality recommendations can lead to user preference amplification, therefore, decreasing the users' exposure to diverse content.
- **Feedback Loop:** User tends to follow the recommendations and the recommendations become the user's interests themselves, leading to amplified biases. Moreover, in the long run, and with repetition of this loop, the amplified biases become the ground truth as the user logs are utilized to train other models [73].

Through the CRS lens: Similar to the Popularity bias, in a conversational setting, users' ability to navigate the recommendations after receiving them gives them the option to reduce the attribute biases in the final list of recommendations as well.

With the inclusion of additional data such as conversations in user interactions, the model's tendency to overemphasize the previous dialogues may increase, potentially exacerbating personalization bias. Therefore, the presence of diverse data sources should be carefully managed to strike a balance between personalization and diversity in recommendations.

5.7. Context Bias

In a work by Zheng et al. [74], the concept of "context bias" is explored as a comprehensive framework for analyzing a collection of recommended items. The study highlights that users' decision-making processes can be influenced by a combination of biases when presented with options that possess different attributes and potentially unique biases for each item. For instance, when browsing a news website, users may encounter various forms of content, such as text and video (modality bias), alongside popularity-driven recommendations (popularity bias). Understanding how this set of biases, collectively referred to as context bias, impacts decision-making requires a broader and more in-depth investigation.

Through the CRS lens: In conversational settings, context bias can play a vital role as the conversational nature of the system makes it more complicated and the dynamics of the model and different existing biases should also be addressed in the same context. In a conversation, the concept of context assumes a broader and more intricate definition compared to its application in traditional recommender systems confined to lists of items. Therefore, the potential exacerbation of context bias becomes a relevant consideration when applied to CRSs.

6. Biases of Attribute-aware Conversational Recommender Systems

Conversational recommender systems introduce a new set of biases that can impact recommendations and user experiences. While traditional biases in recommender systems have been extensively studied, the conversational nature of these systems introduces new biases in distinct ways. In this section, we conduct investigations on the existing biases of conversational recommender systems that do not have natural language understanding and interact with the user based on inquiring about attributes that they want to make decisions upon, such as [62, 1, 75, 76, 77].

6.1. Anchoring Bias

In a conversational setting, the model has the option to utilize the user information throughout the conversation, even through different sessions depending on the system. Therefore, it is a challenge to how this information and the user's behavior and feedback to different recommendations should be utilized. Anchoring bias, or User History bias, happens when the previous recommendations and the user dialogue history in the conversation can anchor subsequent recommendations, leading the system to focus on a particular subset of items and potentially overlooking other options, therefore it is vital to catch the dynamics of user profile while being able to make use of the information. There have been some studies related to this bias. It is investigated that the dynamic nature of conversational systems can amplify the impact of anchoring bias [78, 4]. Also, Ren et al. [5] investigate the differences between old and new user preferences and the way they change.

6.2. Attribute Selection Bias

Attribute selection bias, which can also be called User Preference Assumption bias, refers to a phenomenon in conversational recommender systems where the system becomes biased towards the attributes that users prioritize when making decisions [20]. In each step of the recommendation process, the system may assume that the user wants to base their choices on specific attributes, thereby influencing the recommendations accordingly. This bias can impact the diversity and fairness of the recommendations by potentially overlooking alternative attributes that users might value but are not explicitly expressed. By primarily focusing on a subset of attributes, the system may limit the scope of recommendations, potentially hindering serendipitous discoveries and failing to provide a comprehensive and personalized experience.

6.3. Human-AI Interaction Bias

AI-conversation bias refers to a type of bias where users alter their conversational behavior and speech patterns when interacting with a conversational recommender system that they know is powered by an AI bot [79]. When users are aware that they are conversing with an artificial intelligence rather than a human, they may consciously or unconsciously modify their language, tone, or style of communication. This bias can arise from various factors, including a perceived need to simplify language, adapt to the system's limitations, or conform to social norms associated with human-AI interactions. As a result, the quality and naturalness of the conversation may be affected, potentially leading to a less engaging and authentic user experience [80].

6.4. Modality Bias

Modality bias in a conversational recommender system setting refers to the tendency of multi-modal text generation models, such as the multi-modal GPT-4 model [81], to heavily rely on textual input while paying less attention to non-textual signals, such as visual cues or signals [82]. This bias can limit the system's ability to effectively incorporate and leverage non-textual signals, leading to a potential loss of valuable information and a less holistic understanding of user preferences. Addressing modality bias involves developing more balanced and comprehensive models that can effectively capture and utilize both textual and non-textual cues.

7. Biases of Topic-guided Conversational Recommender Systems

In this section, our focus is to investigate the biases of natural language understanding models and their potential implications when integrated into conversational recommender systems with natural language modules. These types of models can be considered as a general type of the existing CRSs in the literature [8, 83].

7.1. Defective Queries Bias

Defective query bias in a conversational recommender system setting occurs when users intentionally manipulate the system or unknowingly express ambiguous statements, making it difficult for the system to comprehend the conversation and generate useful recommendations [3, 84, 85, 86, 87, 88]. This poses a challenge as the system struggles to fully grasp user intent and preferences. In such situations, the system may need to ask clarifying questions to obtain additional context and disambiguate user queries [89, 90]. However, if the system fails to address this bias effectively, it may lead to the recommendation of low-quality items that do not align with the user's preferences. Consequently, users may need to provide feedback or criticize the initial recommendations to prompt the system to refine its list of suggested items [46, 7, 91]. Addressing Defective Query bias requires the development of robust and adaptive conversational recommender systems that can handle uncertainties, disambiguate user queries, and incorporate user feedback to improve the quality and relevance of recommendations.

7.2. Cognitive Biases

Cognitive bias in language models within a conversational recommender system setting involves the evaluation of large language models in relation to the cognitive biases observed in humans. These biases have the potential to directly impact the system's natural language generation module, influencing it to make irrational decisions depending on whether or not it is affected by these cognitive biases. Detecting and understanding the presence of cognitive biases in language models is crucial to ensure that the recommendations provided are fair and unbiased. By addressing and mitigating these biases, conversational recommender systems can strive to deliver more objective and rational recommendations that are not influenced by human cognitive biases [92].
Similar Notions:

- **Framing bias:** Framing in conversational recommender systems refers to how information presentation influences user perception and decision-making. By addressing framing biases through transparent and balanced recommendations, systems can enhance fairness and effectiveness [93, 94].
- **Uncertainty-aversion bias:** This bias arises from users' negative inclination toward uncertain recommendations [95]. It can impact the effects of explainability techniques on user behavior. Moreover, it can also be studied on the effects of uncertainty in manually labeling datasets.
- **User Trust Bias:** User Trust Bias It is been studied that having a conversational interface will increase the trust rate of users [96, 6]. Karduni et al. [97] argue that the way the faces are shown in news posts affects how much the users trust the platform. Also, the proactivity of the bot is discussed in the works by Kraus et al. [98, 99], Zhu et al. [100] and Lei et al. [101] and it is verified that it affects human trust.

7.3. Unintended Bias

Unintended bias in a conversational recommender system setting refers to the unintentional influence that certain factors or characteristics may have on the recommendations provided [21]. These biases can arise from societal stereotypes, historical data biases, or implicit associations present in language modeling and recommendation algorithms. Unintended biases can result in unequal treatment or favoritism towards certain groups or preferences, leading to potentially unfair or biased recommendations.

7.4. Persona Bias

There is Persona bias in a dialogue system, and therefore in a topic-guided CRS, refers to the detrimental discrepancies in responses that arise when different personas are adopted. There are a few works in dialogue models and recommendation systems that try to adapt the generation process by conforming to the user's persona [102, 103, 104, 105, 106, 107**?**]. These biases manifest in various ways, such as variations in the level of offensiveness or agreement with harmful statements within the generated responses. The adoption of different demographic personas can lead to unequal or biased treatment of users based on their demographic attributes [108]. With a similar outlook, Melchiorre et al. [18] investigate how different user personalities affect the recommendations that they receive in a music recommender system.

Bias Name	Rises From	First Affects
Popularity	RS	DMS
User Log	CRS	Logs
Recomm. Evaluation	RS Training	RS
Attribute	RS	DMS
Position	DMS	NLG
Personalization	RS	DMS
Context	RS	User
Anchoring	DMS	RS
Attribute Selection	DMS	NLG
Human-AI Interaction	User	NLU
Modality	NLU	DMS
Defective Queries	User	NLU
Cognitive	NLG	User
Unintended	NLG (NLU)	User (DMS)
Persona	NLU (NLG)	DMS (DMS)

Table 1
The various biases and their origins are listed, along with their effects on different modules of the CRS. Each of the biases can be traced back to a route in the architecture of a CRS illustrated in Figure 1. The start of route would be where the bias rises from and the initial effecting point of the bias would be the end of the route. This would enable us to navigate through the system in order to investigate the interplay of the different biases and their effects on each other.

8. Limitations and Future Work

When conducting a survey on CRS biases, it's crucial to acknowledge some limitations that might impact the scope and depth of the findings. This paper acknowledges the potential relevance of natural language processing conferences as valuable sources of related references on biases [109]. However, these conferences were not extensively explored here. Additionally, focusing solely on papers published after 2019 might overlook earlier works providing historical context and a deeper understanding of CRS biases' evolution. While the survey aims for wide-ranging coverage, the references might not encompass all relevant literature on each bias. Nevertheless, the survey prioritizes breadth to establish a foundational understanding. Furthermore, lacking formal bias definitions and evaluation metrics may affect results, making their inclusion desirable for better consistency. Lastly, the study's limitations include the absence of experimentation on datasets, models, and presented mitigation methods. Despite these limitations, the survey offers valuable insights into CRS biases, paving the way for more in-depth research and robust approaches to address and mitigate biases in recommendation systems effectively.

In addition to addressing the above limitations, in future research endeavours, an exploration of the intersection between these biases presents a promising avenue. Delving into the intricate interplay of these biases of the system could offer a more comprehensive comprehension. Additionally, leveraging the insights from Table 1, a deeper understanding of how certain biases might synergize within specific system architectures could shed light on the propagation of bias and its effects on user interactions and decision-making processes.

9. Conclusion

This survey paper establishes a non-linear taxonomy of biases in conversational recommender systems. The investigation first covers biases in classic recommender systems and their adaptability to conversational recommender systems, highlighting their nature in conversational settings. Subsequently, biases in conversational recommender systems are explored in two parts: attribute-aware systems and topic-guided systems. By providing this taxonomy, the paper aims to aid researchers and developers in effectively addressing biases, ensuring fairness, transparency, and user trust in these systems' direct impact on user decision-making.

Acknowledgments

This research was partially supported by the Canada CIFAR AI Chair program (Mila), the Facebook research award and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] W. Lei, G. Zhang, X. He, Y. Miao, X. Wang, L. Chen, T.-S. Chua, Interactive path reasoning on graph for conversational recommendation, in: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, 2020, pp. 2073–2083.

[2] Y. Deng, W. Zhang, W. Lam, H. Cheng, H. Meng, User satisfaction estimation with sequential dialogue act modeling in goal-oriented conversational systems, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 2998–3008.

[3] S. Li, B. P. Majumder, J. McAuley, Self-supervised bot play for conversational recommendation with justifications, arXiv preprint arXiv:2112.05197 (2021).

[4] S. Li, R. Xie, Y. Zhu, X. Ao, F. Zhuang, Q. He, User-centric conversational recommendation with multi-aspect user modeling, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 223–233.

[5] Z. Ren, Z. Tian, D. Li, P. Ren, L. Yang, X. Xin, H. Liang, M. de Rijke, Z. Chen, Variational reasoning about user preferences for conversational recommendation, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 165–175.

[6] M. Radensky, J. A. Séguin, J. S. Lim, K. Olson, R. Geiger, "i think you might like this": Exploring effects of confidence signal patterns on trust in and reliance on conversational recommender systems, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 792–804.

[7] H. Li, S. Sanner, K. Luo, G. Wu, A ranking optimization approach to latent linear critiquing for conversational recommender systems, in: Proceedings of the 14th ACM Conference on Recommender Systems, 2020, pp. 13–22.

[8] K. Zhou, W. X. Zhao, S. Bian, Y. Zhou, J.-R. Wen, J. Yu, Improving conversational recommender systems via knowledge graph based semantic fusion, in: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, 2020, pp. 1006–1014.

[9] D. Jannach, L. Chen, Conversational recommendation: A grand ai challenge, 2022. arXiv:2203.09126.

[10] P. P. Ray, Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet of Things and Cyber-Physical Systems (2023).

[11] S. Dai, N. Shao, H. Zhao, W. Yu, Z. Si, C. Xu, Z. Sun, X. Zhang, J. Xu, Uncovering chatgpt's capabilities in recommender systems, 2023. arXiv:2305.02182.

[12] D. Jin, L. Wang, H. Zhang, Y. Zheng, W. Ding, F. Xia, S. Pan, A survey on fairness-aware recommender systems, 2023. arXiv:2306.00403.

[13] S. Milano, M. Taddeo, L. Floridi, Recommender systems and their ethical challenges, Ai & Society 35 (2020) 957–967.

[14] S. Gupta, H. Wang, Z. Lipton, Y. Wang, Correcting exposure bias for link recommendation, in: International Conference on Machine Learning, PMLR, 2021, pp. 3953–3963.

[15] H. Liu, D. Tang, J. Yang, X. Zhao, H. Liu, J. Tang, Y. Cheng, Rating distribution calibration for selection bias mitigation in recommendations, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 2048–2057.

[16] C. Lin, X. Liu, G. Xv, H. Li, Mitigating sentiment bias for recommender systems, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 31–40.

[17] Z. Zhu, Y. He, X. Zhao, Y. Zhang, J. Wang, J. Caverlee, Popularity-opportunity bias in collaborative filtering, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021, pp. 85–93.

[18] A. B. Melchiorre, E. Zangerle, M. Schedl, Personality bias of music recommendation algorithms, in: Proceedings of the 14th ACM Conference on Recommender Systems, 2020, pp. 533–538.

[19] D. Jannach, A. Manzoor, W. Cai, L. Chen, A survey on conversational recommender systems, ACM Computing Surveys 54 (2021) 1–36. URL: https://doi.org/10.1145%2F3453154. doi:10.1145/3453154.

[20] S. Lin, Z. Zhu, J. Wang, J. Caverlee, Towards fair conversational recommender systems, arXiv preprint arXiv:2208.03854 (2022).

[21] T. Shen, J. Li, M. R. Bouadjenek, Z. Mai, S. Sanner, Unintended bias in language model-driven conversational recommendation, arXiv preprint arXiv:2201.06224 (2022).

[22] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman,

A. Galstyan, A survey on bias and fairness in machine learning, 2022. arXiv:1908.09635.

[23] S. L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (technology) is power: A critical survey of" bias" in nlp, arXiv preprint arXiv:2005.14050 (2020).

[24] Y. Wang, W. Ma, M. Zhang, Y. Liu, S. Ma, A survey on the fairness of recommender systems, ACM Transactions on Information Systems 41 (2023) 1–43.

[25] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, X. He, Bias and debias in recommender system: A survey and future directions, 2021. arXiv:2010.03240.

[26] A. Lin, J. Wang, Z. Zhu, J. Caverlee, Quantifying and mitigating popularity bias in conversational recommender systems, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 1238–1247.

[27] D. Jannach, Evaluating conversational recommender systems, Artificial Intelligence Review 56 (2022) 2365–2400. URL: https://doi.org/10.1007%2Fs10462-022-10229-x. doi:10.1007/s10462-022-10229-x.

[28] W. Brown, A. Agarwal, Diversified recommendations for agents with adaptive preferences, Advances in Neural Information Processing Systems 35 (2022) 26066–26077.

[29] A. Zhang, J. Zheng, X. Wang, Y. Yuan, T.-S. Chua, Invariant collaborative filtering to popularity distribution shift, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 1240–1251.

[30] H. Abdollahpouri, M. Mansoury, R. Burke, B. Mobasher, The connection between popularity bias, calibration, and fairness in recommendation, in: Proceedings of the 14th ACM Conference on Recommender Systems, 2020, pp. 726–731.

[31] H. Abdollahpouri, R. Burke, B. Mobasher, Managing popularity bias in recommender systems with personalized re-ranking, arXiv preprint arXiv:1901.07555 (2019).

[32] L. Michiels, J. Leysen, A. Smets, B. Goethals, What are filter bubbles really? a review of the conceptual and empirical work, in: Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, 2022, pp. 274–279.

[33] W. Wang, F. Feng, L. Nie, T.-S. Chua, User-controllable recommendation against filter bubbles, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1251–1261.

[34] S. Frumerman, G. Shani, B. Shapira, O. Sar Shalom, Are all rejected recommendations equally bad? towards analysing rejected recommendations, in: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, 2019, pp. 157–165.

[35] Z. Nazari, P. Chandar, G. Fazelnia, C. M. Edwards, B. Carterette, M. Lalmas, Choice of implicit signal matters: Accounting for user aspirations in podcast recommendations, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 2433–2441.

[36] Z. Xu, P. Wei, W. Zhang, S. Liu, L. Wang, B. Zheng, Ukd: Debiasing conversion rate estimation via uncertainty-regularized knowledge distillation, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 2078–2087.

[37] X. Zhang, S. Dai, J. Xu, Z. Dong, Q. Dai, J.-R. Wen, Counteracting user attention bias in music streaming recommendation via reward modification, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 2504–2514.

[38] Y. Zheng, C. Gao, X. Li, X. He, Y. Li, D. Jin, Disentangling user interest and conformity for recommendation with causal embedding, in: Proceedings of the Web Conference 2021, 2021, pp. 2980–2991.

[39] A. Dash, A. Chakraborty, S. Ghosh, A. Mukherjee, K. P. Gummadi, When the umpire is also a player: Bias in private label product recommendations on e-commerce marketplaces, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 873–884.

[40] J. McInerney, B. Brost, P. Chandar, R. Mehrotra, B. Carterette, Counterfactual evaluation of slate recommendations with sequential reward interactions, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1779–1788.

[41] W. Zhang, W. Bao, X.-Y. Liu, K. Yang, Q. Lin, H. Wen, R. Ramezani, Large-scale causal approaches to debiasing post-click conversion rate estimation with multi-task learning, in: Proceedings of The Web Conference 2020, 2020, pp. 2775–2781.

[42] Z. Ovaisi, R. Ahsan, Y. Zhang, K. Vasilaky, E. Zheleva, Correcting for selection bias in learning-to-rank systems, in: Proceedings of The Web Conference 2020, 2020, pp. 1863–1873.

[43] J. Huang, H. Oosterhuis, M. De Rijke, H. Van Hoof, Keeping dataset biases out of the simulation: A debiased simulator for reinforcement learning based recommender systems, in: Proceedings of the 14th ACM Conference on Recommender Systems, 2020, pp. 190–199.

[44] T. Yang, C. Luo, H. Lu, P. Gupta, B. Yin, Q. Ai, Can clicks be both labels and features? unbiased behavior feature collection and uncertainty-aware

learning to rank, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 6–17.

[45] I. Szpektor, D. Cohen, G. Elidan, M. Fink, A. Hassidim, O. Keller, S. Kulkarni, E. Ofek, S. Pudinsky, A. Revach, et al., Dynamic composition for conversational domain exploration, in: Proceedings of The Web Conference 2020, 2020, pp. 872–883.

[46] T. Yu, Y. Shen, H. Jin, A visual dialog augmented interactive recommender system, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 157–165.

[47] R. Y. Pang, J. Cenatempo, F. Graham, B. Kuehn, M. Whisenant, P. Botchway, K. Stone Perez, A. Koenecke, Auditing cross-cultural consistency of human-annotated labels for recommendation systems, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 1531–1552.

[48] A. Raul, A. P. Dharwadker, B. Schumitsch, Cam2: Conformity-aware multi-task ranking model for large-scale recommender systems, arXiv preprint arXiv:2304.08562 (2023).

[49] Y. Yang, C. Huang, L. Xia, C. Huang, D. Luo, K. Lin, Debiased contrastive learning for sequential recommendation, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 1063–1073.

[50] A. D'Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, Y. Halpern, Fairness is not static: deeper understanding of long term fairness via simulation studies, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 525–534.

[51] H. Wen, X. Yi, T. Yao, J. Tang, L. Hong, E. H. Chi, Distributionally-robust recommendations for improving worst-case user experience, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 3606–3610.

[52] M. Mladenov, E. Creager, O. Ben-Porat, K. Swersky, R. Zemel, C. Boutilier, Optimizing long-term social welfare in recommender systems: A constrained matching approach, in: International Conference on Machine Learning, PMLR, 2020, pp. 6987–6998.

[53] G. Hiranandani, W. Vijitbenjaronk, S. Koyejo, P. Jain, Optimization and analysis of the pap@ k metric for recommender systems, in: International Conference on Machine Learning, PMLR, 2020, pp. 4260–4270.

[54] K. Christakopoulou, M. Traverse, T. Potter, E. Marriott, D. Li, C. Haulk, E. H. Chi, M. Chen, Deconfounding user satisfaction estimation from response rate bias, in: Proceedings of the 14th ACM Conference on Recommender Systems, 2020, pp. 450–455.

[55] F. Diaz, A. Ferraro, Offline retrieval evaluation without evaluation metrics, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 599–609.

[56] A. Raj, M. D. Ekstrand, Measuring fairness in ranked results: An analytical and empirical comparison, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 726–736.

[57] D. Cohen, B. Mitra, O. Lesota, N. Rekabsaz, C. Eickhoff, Not all relevance scores are equal: Efficient uncertainty and calibration modeling for deep retrieval models, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 654–664.

[58] T. De Jonge, D. Hiemstra, Unfair: Search engine manipulation, undetectable by amortized inequity, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 830–839.

[59] X. Wang, W. Zhu, C. Liu, Social recommendation with optimal limited attention, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 1518–1527.

[60] A. Ghandeharioun, J. H. Shen, N. Jaques, C. Ferguson, N. Jones, A. Lapedriza, R. Picard, Approximating interactive human evaluation with self-play for open-domain dialog systems, Advances in Neural Information Processing Systems 32 (2019).

[61] W. Sun, S. Zhang, K. Balog, Z. Ren, P. Ren, Z. Chen, M. de Rijke, Simulating user satisfaction for the evaluation of task-oriented dialogue systems, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2499–2506.

[62] X. Wang, K. Zhou, J.-R. Wen, W. X. Zhao, Towards unified conversational recommender systems via knowledge-enhanced prompt learning, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 1929–1937.

[63] R. Islam, K. N. Keya, S. Pan, J. Foulds, Mitigating demographic biases in social media-based recommender systems, KDD (Social Impact Track) (2019).

[64] Y. Ying, F. Zhuang, Y. Zhu, D. Wang, H. Zheng, Camus: Attribute-aware counterfactual augmentation for minority users in recommendation, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 1396–1404.

[65] F. Eskandanian, N. Sonboli, B. Mobasher, Power

of the few: Analyzing the impact of influential users in collaborative recommender systems, in: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, 2019, pp. 225–233.

[66] Y. Li, H. Chen, Z. Fu, Y. Ge, Y. Zhang, User-oriented fairness in recommendation, in: Proceedings of the Web Conference 2021, 2021, pp. 624–632.

[67] D. Harwell, The Accent Gap, https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/, 2018.

[68] M. Cogswell, J. Lu, R. Jain, S. Lee, D. Parikh, D. Batra, Dialog without dialog data: Learning visual dialog agents from vqa data, Advances in Neural Information Processing Systems 33 (2020) 19988–19999.

[69] Z. Zheng, Z. Qiu, H. Xiong, X. Wu, T. Xu, E. Chen, X. Zhao, Ddr: Dialogue based doctor recommendation for online medical service, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 4592–4600.

[70] M. Ruffini, V. Bellini, A. Buchholz, G. Di Benedetto, Y. Stein, Modeling position bias ranking for streaming media services, in: Companion Proceedings of the Web Conference 2022, 2022, pp. 72–76.

[71] C. Zhu, Z. Yang, R. Gmyr, M. Zeng, X. Huang, Leveraging lead bias for zero-shot abstractive news summarization, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1462–1471.

[72] D. Kalimeris, S. Bhagat, S. Kalyanaraman, U. Weinsberg, Preference amplification in recommender systems, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 805–815.

[73] M. Mansoury, H. Abdollahpouri, M. Pechenizkiy, B. Mobasher, R. Burke, Feedback loop and bias amplification in recommender systems, in: Proceedings of the 29th ACM international conference on information & knowledge management, 2020, pp. 2145–2148.

[74] Z. Zheng, Z. Qiu, T. Xu, X. Wu, X. Zhao, E. Chen, H. Xiong, Cbr: context bias aware recommendation for debiasing user modeling and click prediction, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 2268–2276.

[75] S. Li, W. Lei, Q. Wu, X. He, P. Jiang, T.-S. Chua, Seamlessly unifying attributes and items: Conversational recommendation for cold-start users, ACM Transactions on Information Systems (TOIS) 39 (2021) 1–29.

[76] X. Ren, H. Yin, T. Chen, H. Wang, Z. Huang, K. Zheng, Learning to ask appropriate questions in conversational recommendation, in: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, 2021, pp. 808–817.

[77] K. Xu, J. Yang, J. Xu, S. Gao, J. Guo, J.-R. Wen, Adapting user preference to online feedback in multi-round conversational recommendation, in: Proceedings of the 14th ACM international conference on web search and data mining, 2021, pp. 364–372.

[78] C. Qu, L. Yang, M. Qiu, W. B. Croft, Y. Zhang, M. Iyyer, Bert with history answer embedding for conversational question answering, in: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, 2019, pp. 1133–1136.

[79] A. Følstad, T. Araujo, E. L.-C. Law, P. B. Brandtzaeg, S. Papadopoulos, L. Reis, M. Baez, G. Laban, P. McAllister, C. Ischen, et al., Future directions for chatbot research: an interdisciplinary research agenda, Computing 103 (2021) 2915–2942.

[80] E. Hosseini-Asl, B. McCann, C.-S. Wu, S. Yavuz, R. Socher, A simple language model for task-oriented dialogue, Advances in Neural Information Processing Systems 33 (2020) 20179–20191.

[81] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.

[82] Z. Tian, Z. Xie, F. Lin, Y. Song, A multi-view meta-learning approach for multi-modal response generation, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 1938–1947.

[83] R. Sarkar, K. Goswami, M. Arcan, J. P. McCrae, Suggest me a movie for tonight: Leveraging knowledge graphs for conversational recommendation, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 4179–4189.

[84] Z. Lu, R. H. Kazi, L.-y. Wei, M. Dontcheva, K. Karahalios, Streamsketch: Exploring multi-modal interactions in creative live streams, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–26.

[85] T. Shen, Z. Mai, G. Wu, S. Sanner, Distributional contrastive embedding for clarification-based conversational critiquing, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 2422–2432.

[86] Z. Wang, Y. Tu, C. Rosset, N. Craswell, M. Wu, Q. Ai, Zero-shot clarifying question generation for conversational search, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 3288–3298.

[87] J. Liao, X. Zhao, J. Zheng, X. Li, F. Cai, J. Tang, Ptau: Prompt tuning for attributing unanswerable questions, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp.

1219–1229.

[88] M. S. Mirzaei, K. Meshgi, S. Sekine, Is this question real? dataset collection on perceived intentions and implicit attack detection, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 2850–2859.

[89] A. Montazeralghaem, J. Allan, P. S. Thomas, Large-scale interactive conversational recommendation system using actor-critic framework, in: Proceedings of the 15th ACM Conference on Recommender Systems, 2021, pp. 220–229.

[90] M. Aliannejadi, H. Zamani, F. Crestani, W. B. Croft, Asking clarifying questions in open-domain information-seeking conversations, in: Proceedings of the 42nd international acm sigir conference on research and development in information retrieval, 2019, pp. 475–484.

[91] P. Yang, H. Huang, W. Wei, X.-L. Mao, Toward real-life dialogue state tracking involving negative feedback utterances, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 2222–2232.

[92] E. Jones, J. Steinhardt, Capturing failures of large language models via human cognitive biases, Advances in Neural Information Processing Systems 35 (2022) 11785–11799.

[93] M. Reiter-Haas, Exploration of framing biases in polarized online content consumption, in: Companion Proceedings of the ACM Web Conference 2023, 2023, pp. 560–564.

[94] M. Mulder, O. Inel, J. Oosterman, N. Tintarev, Operationalizing framing to support multiperspective recommendations of opinion pieces, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 478–488.

[95] S. Park, J. Y. Park, H. Chin, J.-h. Kang, M. Cha, An experimental study to understand user experience and perception bias occurred by fact-checking messages, in: Proceedings of the Web Conference 2021, 2021, pp. 2769–2780.

[96] A. Gupta, D. Basu, R. Ghantasala, S. Qiu, U. Gadiraju, To trust or not to trust: How a conversational interface affects trust in a decision support system, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 3531–3540.

[97] A. Karduni, R. Wesslen, D. Markant, W. Dou, Images, emotions, and credibility: Effect of emotional facial images on perceptions of news content bias and source credibility in social media, arXiv preprint arXiv:2102.13167 (2021).

[98] M. Kraus, N. Wagner, N. Untereiner, W. Minker, Including social expectations for trustworthy proactive human-robot dialogue, in: Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, 2022, pp. 23–33.

[99] M. Kraus, N. Wagner, W. Minker, Effects of proactive dialogue strategies on human-computer trust, in: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, 2020, pp. 107–116.

[100] Y. Zhu, J.-Y. Nie, K. Zhou, P. Du, H. Jiang, Z. Dou, Proactive retrieval-based chatbots based on relevant knowledge and goals, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2000–2004.

[101] W. Lei, Y. Zhang, F. Song, H. Liang, J. Mao, J. Lv, Z. Yang, T.-S. Chua, Interacting with non-cooperative user: A new paradigm for proactive dialogue policy, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 212–222.

[102] S. Fatahi, M. Mousavifar, J. Vassileva, Investigating the effectiveness of persuasive justification messages in fair music recommender systems for users with different personality traits, in: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization, 2023, pp. 66–77.

[103] Z. Han, S. Zhang, X. Zhang, Persona consistent dialogue generation via contrastive learning, in: Companion Proceedings of the ACM Web Conference 2023, 2023, pp. 196–199.

[104] C. Xu, P. Li, W. Wang, H. Yang, S. Wang, C. Xiao, Cosplay: Concept set guided personalized dialogue generation across both party personas, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 201–211.

[105] J.-C. Gu, H. Liu, Z.-H. Ling, Q. Liu, Z. Chen, X. Zhu, Partner matters! an empirical study on fusing personas for personalized response selection in retrieval-based chatbots, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 565–574.

[106] Z. Ma, Z. Dou, Y. Zhu, H. Zhong, J.-R. Wen, One chatbot per person: Creating personalized chatbots based on implicit user profiles, in: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, 2021, pp. 555–564.

[107] A. Tigunova, A. Yates, P. Mirza, G. Weikum, Listening between the lines: Learning personal attributes from conversations, in: The World Wide Web Conference, 2019, pp. 1818–1828.

[108] E. Sheng, J. Arnold, Z. Yu, K.-W. Chang, N. Peng, Revealing persona biases in dialogue systems, arXiv preprint arXiv:2104.08728 (2021).

[109] Y. Wan, W. Wang, P. He, J. Gu, H. Bai, M. Lyu,

Biasasker: Measuring the bias in conversational ai system, 2023. arXiv:2305.12434.