Causal Inference for Recommendation: Foundations, Methods and Applications

Shuyuan Xu, Jianchao Ji, Yunqi Li, Yingqiang Ge, Juntao Tan, Yongfeng Zhang

Abstract—Recommender systems are important and powerful tools for various personalized services. Traditionally, these systems use data mining and machine learning techniques to make recommendations based on correlations found in the data. However, relying solely on correlation without considering the underlying causal mechanism may lead to various practical issues such as fairness, explainability, robustness, bias, echo chamber and controllability problems. Therefore, researchers in related area have begun incorporating causality into recommendation systems to address these issues. In this survey, we review the existing literature on causal inference in recommender systems. We discuss the fundamental concepts of both recommender systems and causal inference as well as their relationship, and review the existing work on causal methods for different problems in recommender systems. Finally, we discuss open problems and future directions in the field of causal inference for recommendations.

Index Terms—Recommender Systems, Causal Inference

1 Introduction

Recommender systems have been recognized as one of the most effective tools to alleviate the information overloading, and have been widely deployed in many real-world systems, such as e-commerce platforms (e.g., Amazon, eBay), social networks (e.g., Facebook, Twitter), video-sharing platforms (e.g., Youtube, TikTok) and streaming services (e.g., Netflix, Hulu). In general, these systems use advanced techniques to learn users' preferences from historical data, along with collected user, item, and content information. And the development of these techniques has advanced rapidly in recent years.

Generally speaking, recommendation algorithms can be categorized into three major types: collaborative filtering, content-based recommendation and hybrid methods [1, 2, 3]. Collaborative filtering (CF) models are based on a key idea that similar users may share similar interest and similar items may be liked by similar users. Early memory-based CF models, such as user-based CF [4, 5] and item-based CF [6, 7], take the row or column vectors of the user-item rating matrix as the user and item vector representations, and calculate the similarity between users or items for recommendation based on pre-defined similarity functions such as cosine similarity and Pearson correlation coefficient. To extract latent semantic meanings from the matrix, researchers later explored learned user and item vector representations. This started with Latent Factor Models (LFM) such as matrix factorization [8], probabilistic matrix factorization [9] and factorization machines [10], which are widely adopted models in practice. In these models, each user and item is learned as a latent representation to calculate the matching score of each user-item pair, usually based on inner-product. The development of deep learning and neural networks has further extended CF models. For example, [11, 12, 13, 14] adopts simple user and item representations (e.g., one-hot vectors) and learns a complex matching function. [15, 16, 17, 18, 19] learn complex user and item representations and adopt a simple matching function (e.g., inner product). User representations can also be directly calculated from historical interactions, such as in sequential recommendation [20, 21]. Content-based recommendation will utilize rich information about users and items, or even context information, to enhance recommendation. In order to learn the similarities among items based on the side information, the representation approaches applied by the content-based recommendations have been developed from simple models such as TF-IDF [22] to deep learning based models such as DNN [23], CNN [24], etc. Hybrid approaches combine collaborative filtering and content-based methods, which exploit the benefit of both methods and avoid their certain limitations [1, 2, 25].

The foundation of traditional recommendation algorithms is mining or learning the correlative pattern from data. For example, many collaborative filtering models aim to learn the user-item correlative pattern, some content-based recommendation models aim to learn the feature-feature correlative pattern. However, the real-world applications are driven by underlying causal mechanisms, pure correlative learning without considering the causation will lead to some practical issues. We take the classic "beer and diapers" problems as an example. Pure correlative learning will learn the strong correlation pattern between beer and diapers, thus recommend beer for customers bought diapers or vice versa. However, the underlying mechanism is that young fathers usually buy beer and diapers together, and recommending beer or diapers without considering the underlying mechanism will cause confusion and further hurt user's satisfaction. Therefore, it is important to advance from correlative learning to causal learning.

Formally speaking, causal inference studies the causal relation between cause and effect, where cause takes responsible for the effect. Two famous and popular frameworks are the potential outcome framework (also known as the Neyman–Rubin Potential Outcomes or the Rubin Causal Model) [26] and the structural causal model (SCM) [27, 28].

The authors are with the Department of Computer Science, Rutgers University, USA.
Emails: {shuyuan.xu, jianchao.ji, yunqi.li, yingqiang.ge, juntao.tan, yongfeng.zhang}@rutgers.edu

Both causal frameworks contribute to the development of causal recommendations. By leveraging the underlying causal mechanisms in recommender systems, causal recommendation is able to handle different practical issues, including explainability, fairness, robustness, uplift, and unbiasedness.

Contribution of this survey. In this survey, we aim to provide a comprehensive review of causal inference for recommendation. We first introduce the fundamental knowledge of recommender systems and then discuss existing work of causal inference for recommendation. Specifically, we explore the causal inference in recommender systems in two dimensions. The first dimension follows the pipeline of causal inference, including concepts, notations, and techniques in causal inference, and the connection between causal inference and recommender systems. The second dimension follows the practical problems in recommendation, including problem introduction, causal methods, and open problems. More specifically, we include explainability, fairness, robustness, uplift-based, unbiasedness in recommendation. Finally, we highlight several open problems in causal inference for recommendation that remain to be addressed.

Difference with Existing Surveys. Several surveys in recommender systems or causal inference have been published in recent years. For example, Zhang et al. [29] and Chen et al. [30] review explainable recommendation, Li et al. [31] and Wang et al. [32] review fairness in recommendation, Ge et al. [33], Wang et al. [34] and Fan et al. [35] summarize trustworthy recommender systems, Chen et al. [36] review bias in recommendation, Zhang et al. [2] review the deep learning based recommendation algorithms, Ko et al. [37] provide a comprehensive review of recommender systems, Yao et al. [38] provide a comprehensive review of causal inference methods, Guo et al. [39] and Vowels et al. [40] summarize existing methods on causal structural learning and causal discovery. Gao et al. [41] summarize existing work on causal inference in recommender systems. Unlike Gao et al. [41] mainly introduce existing work in perspective of recommender systems, our survey provide systematic review in perspective of both causal inference and recommender systems.

Organization. This survey is organized as follows: Section 2 introduces the preliminaries of recommender systems. From Section 3 to 7, we introduce fundamental knowledge of causal inference and the connection with recommender systems. Section 8 to 12 introduce existing causal methods on explainable recommendation, fairness in recommendation, uplift-based recommendation, robust recommendation, unbiased recommendation, respective. In Section 13, we discuss some open problems and future directions in causal inference for recommendation. Section 14 concludes this survey.

2 Preliminaries for Recommender Systems

In general, recommender systems aim to model user preferences based on collected information, including user profile, item profile, and user-item interactions, and further predict users' future interactions. User profile represents the registered information of the user, which may include user id, user age, user gender, user income, etc. Recommender systems may only use partial information for recommendation (e.g., using user id only). The term "items" represents different objects in differnt recommender systems (e.g., product in e-commerce, other users in social networks, videos in online video platform, etc.). According to different definition of "item", item profile may include different item features. For example, products in e-commerce may take brand, category, price, image , etc. in item profile; videos in online video platform item profile may take video length, content description, etc. in video recommendation; other users in social networks may take corresponding user profiles as item profile. Similarly, recommender systems may only partial information of item profile for recommendation. Interactions refer to possible user behaviors towards items according to defined task (e.g., click, purchase, rate, add-to-cart, review for e-commerce recommendation, like, dislike, share for video recommendation, etc.). In general recommender systems, interactions are typically represented in two ways, one is explicit feedback, the other is implicit feedback. Explicit feedback, such as ratings and reviews, is the explicit representation of users' preference (e.g., rating score as 5 means that user like this item), while implicit feedback, such as click, is collected during user-system interaction process and implicitly represent users' preference (e.g., user's click behavior means that it is likely that user likes the corresponding item).

Traditional recommendation algorithms can be roughly categorized into collaborative filtering, content-based recommendation and hybrid models. The basic idea of collaborative filtering (CF) is that similar users may share similar interests and similar items may be likede by similar users. CF methods can be further divided into memory-based CF and model-based CF. memory-based CF makes predictions by a simple similarity measurement over historical data. For example, user-based CF [4, 5] or item-based CF [6, 7] takes the row or column vector of the user-item rating matrix as the representation of each user or item and calculate the similarity by a simple measurement such as cosine similarity. Model-based CF leverage a model to learn the representation of users and items to make predictions. It starts from Latent Factor Models, such as matrix factorization [8], probabilistic matrix factorization [9], tensor factorization [42], etc. Deep learning and neural networks have further extend CF models. Deep CF methods can be further divided into similarity learning approach and representation learning approach. The similarity learning approaches [15, 16, 17, 18, 19] leverage simple representation of users and items (e.g., one-hot vectors) and learns a complex matching function to make prediction on each user-item pair. The representation learning approaches learn complex representation of users and items, and then apply a simple matching function (e.g., inner product) to calculate the prediction scores. Content-based recommendation [23, 24, 43, 44, 45, 46, 47], on the other hand, replies on rich user and item profile to recommend items similar to the ones the user prefered in the past. For example, in a movie recommender system, the model tries to understand the features (e.g., actors, directors, genres, tags, etc.) of movies that a user has rate highly in the past. Then, only the movies that match the preferred features of the user would be recommended. Hybrid models combine collaborative filtering and content-based methods, which exploit the benefit of both methods and avoid their certain limitations [1, 2, 25, 48, 49]. Moreover, several works, such as [50, 51], have empirically demonstrated that the hybrid approaches are able to achieve more accurate recommendation than pure collaborative and content-based methods.

Besides above traditional recommendation algorithms, there are some other recommendation algorithms. Sequential recommendation [52] (also related to session-based or session-aware recommendation), which leverage the timestamp information of interactions to suggest items, have become increasingly popular in academic research and industrial application. Traditional sequential recommendation models employ simple machine learning approaches to model sequential data, such as Markov chain [53], session-based KNN [54]. With the development of deep learning techniques, many deep models obtain tremendous achievements in sequential recommendation, including RNN [55], LSTM [56], CNN [57, 58], attention models [59] and memory networks[60]. Moreover, with increasing success achieved by foundation models (e.g., Large Language Models) on natural language tasks (e.g., T5 [61], GPT-3 [62], OPT [63], PaLM [64]), recommender system community, leverage the unique characteristic of recommender systems, has developed the research on personalized foundation models. For example, P5 [65], as a pretrain, personalized prompt,and predict paradigm for recommendation, formulates recommendation as a language understanding and generation task to serve as a foundation model for many recommendation tasks.

The recommendation models learn users' preference based on collected information, and make recommendation based on learned preference. Specifically, a recommender system will provide a personalized recommendation list along with possible explanations to a specific user. Recommender systems will first predict user's preference towards a set of candidate items. Then the system will rank candidate items to provide personalized recommendation list. It is worth mentioning that the ranking process is not necessarily solely based on the predicted scores provided by the recommendation algorithm. It is possible to re-rank the list based on different demands, such as diversity, fairness, some business purpose, etc. After generating personalized recommendation list, some recommendation systems may provide explanations along with recommendations. The explanations can be either generated simultaneously with the recommendation or after the recommendation, depending on the recommendation model is explainable or black-box.

To evaluate the performance of recommender systems, it is important to define the characteristics of a good recommender system and quantify the characteristics. For a recommendation model with ability of predicting rating scores, a excellent model should be able to predict accurate ratings. Therefore, RMSE or MSE is used to evaluate the recommendation performance. By considering the accuracy of ranking list and whether the user's prefered items recommended by the list, some commonly used metrics include Precision, Recall, F-Measure, NDCG, ROC Curve, AUC, MRR, etc. Besides above metrics used to evaluate recommendation performance, some metrics are used to evaluate the recommendation model in perspective of other purpose. For example, Absolute Difference (AD) [66] is used to evaluate the fairness of recommender systems.

3 Causal Notations in Recommendation

Causal inference is a critical research topic stemmed from statistics [28, 67, 68], and has been widely used in many domains for decades, such as computer science, public policy, economic, etc. In this section, we introduce causal notations and demonstrate how to apply them in recommendation.

3.1 What is Causation

Causation (also refer to as causality) is a terminology that is usually compared to and discussed with correlation. Although both correlation and causation explore the relationship between variables, it is well known that "correlation does not imply causation" [68]. Causation takes a step further than correlation. Intuitively, causation explicitly applies to the case that event A causes event B. On the other hand, correlation is a much simple relation that event A is related to event B, but one event does not necessarily cause another event to happen. For example, a study has shown that the data of monthly ice cream sales is highly related to the number of monthly shark attacks across the United States. Although the two variables are highly correlated, it is impossible to conclude that consuming ice cream causes shark attacks (or vice versa). It is more likely that both ice cream sales and shark attacks increase in the summer due to other factors such as warm weather, which leads to both variables being correlated. Similar examples can be found in recommendations. The beer-and-diapers story is a good example to illustrate the difference between causation and correlation in recommendation. There is an observation that beer and diapers sell well together. Based on pure correlative learning, beer should be recommended for customers who bought diapers or vice versa because of the strong correlation pattern between beers and diapers. However, the underlying causal mechanism is that young fathers may pick up some diapers while buying beer. Therefore, directly recommending items without considering the underlying causation may lead to confusion and scarified recommendation performance. In general, understanding causation helps us to better understand how the world works and can improve the performance of recommendation systems.

To theoretically study the causation, it is required to understand the mathematical representation of causation. In general, there are two commonly used frameworks for causal inference, one is the potential outcomes framework (also known as the Neyman–Rubin Potential Outcomes or the Rubin Causal Model) [26] and the other is the structural causal model framework [27, 28] proposed by Pearl. Existing works usually introduce two framework separately, however, we think both frameworks are logically equivalent [28] and follow the similar intuition. In the following sections, we will introduce those two frameworks following the intuitive idea of causation, including the connections and differences of two frameworks.

3.2 Key mathematical notations of Causation.

Causal inference refers to a process of drawing a conclusion that a specific *treatment* was the "cause" of the *outcome* that was observed [69]. In this case, the atomic goal is to estimate the outcome if any specific treatment has been applied. Both frameworks use mathematical notations to represent the desired value. For Rubin Causal Model, the basic element is called potential outcome.

Definition 1. ***(Potential Outcome)*** *A potential outcome is the outcome for an individual under a possible treatment*

Let X ($X \in \{x_1, x_2, \cdots, x_n\}$) denotes the treatment, where n is the total number of possible treatments. Most of the literature considers the binary treatment, for example, taking medicine is denoted as $X = 1$ and not taking medicine is denoted as $X = 0$. Under the binary treatment, the group of individuals with treatment $X = 1$ is named as the *treated group*, and the group of individuals with treatment $X = 0$ is called as the *control group*. Generally, the potential outcome of treatment with value x_i is denoted as $Y(X = x_i)$, which can be simplified as $Y(x_i)$. The average potential outcome of treatment with value x_i can be denoted as $\mathbb{E}[Y(x_i)]$. For any individual, only one treatment can be applied while keeping other variables unchanged, thus only one potential outcome can be observed. Therefore, potential outcomes can be further divided into two categories, the observed one is named as observed outcome while the remaining unobserved potential outcomes are named as counterfactual outcomes.

In recommendation, the outcome is usually defined as user behavior (e.g., click, purchase) or user preference (e.g., rating). Unbiased recommendation models define the treatment as exposure, in which the observed feedback Y (i.e., observed outcome) can be modeled as the product of two unobserved variables exposure O and relevance R (i.e., $Y = O \cdot R$) [70, 71, 72, 73, 74]. More specifically, in recommender systems, $Y = 0$ can be either negative samples (i.e., $R = 0$) or potential positive samples (i.e., $R = 1$, $O = 0$), which lead to data bias in recommendation. To achieve personalized recommendation, the models usually estimate the potential outcome $Y_{u,v}$ for a certain user-item pair (u, v) (i.e., $Y_{u,v} = O_{u,v} \cdot R_{u,v}$). By correctly estimating potential outcome $Y_{u,v}(O_{u,v} = 1)$ (i.e., $R_{u,v} = Y_{u,v}(O_{u,v} = 1)$), the designed model is able to achieve unbiased recommendation. Uplift-based recommendation models define the treatment as recommendation (i.e., 1 for recommended, 0 for not recommended) [75]. For each observed user-item pair, only one treatment can be observed (i.e., recommended or not recommended). Therefore, it is a challenge of estimating the counterfactual outcome to calculate the uplift value for recommendation. To achieve fairness, the treatment can also be defined as the sensitive attribute [76](e.g., 1 for privileged group and 0 for disadvantaged group).

Besides the treatment variable and the outcome variable, some other variables can be observed, and they can be further categorized as pre-treatment variables and the post-treatment variables [38].

Definition 2. *(Pre-treatment Variables) Pre-treatment variables are the variables that are not affected by the treatment, which are also named as background variables.*

Definition 3. *(Post-treatment Variables) Post-treatment variables are the variables that are affected by the treatment.*

Different recommendation scenario may include different information and causal mechanisms, thus the specific definition of pre-treatment variables and post-treatment variables may vary.

In addition to the potential outcome, Pearl uses another popular representation which distinguishes correlation and causation using do-operation [27, 68] from the perspective of probability. Supposed that X denotes the treatment and Y denotes the outcome, correlation and causation pursue different probabilities. Specifically, correlation estimates the conditional probability $P(Y|X)$ from observational data to determine the correlative relation between X and Y. By contrast, causal inference estimates $P(Y|do(X = x_i))$ representing the outcome under a possible treatment x_i, where do-operation intuitively denotes applying the treatment instead of observing the treatment. The average outcome of applying treatment x_i can be represented as $\mathbb{E}[Y|do(X = x_i)]$. A specific probability $P(Y = y|do(X = x))$ can be simplified as $P(y|do(x))$. As we mentioned before, existing causal frameworks follow the same intuition, therefore, the mathematical notations of do-operations and potential outcomes can be converted to each other in most cases. For example, in unbiased recommendation model, the treatment is usually defined as exposure. The results for a user-item pair (u, v) under exposure can be expressed as $P(Y|u, v, do(X = 1))$ where Y is the outcome and X is the exposure variable. If we define variable V as exposed items, then it can also be represented by $P(Y|u, do(V = v))$. Similarly, the causal notations in uplift-based recommendation and fairness for recommendation can also be expressed by do-operations.

By defining do-operations, ***intervention*** as a basic concept in causal inference can be formally defined. As we mentioned above, the do-operation denotes applying the treatment, which can be also defined as the *intervention* on the treatment variable. We will introduce more details in section 6. ***Counterfactual*** is an important concepts in both the potential outcome framework and structural causal model, which represents the difference with factual. More specifically, counterfactual represents the scenario that the treatment variable had a different value compared with the observed value in the factual world. For example, considering the treatment as taking drugs and the outcome as recovery, a patient who took drugs and recovered may wonder if he would have been recovered if he hadn't taken the drugs. In this case, in the factual world, the patient took drugs and recovered, and in the counterfactual world, the patient did not take drugs and we wonder if he would recover. Similar example can be observed in recommender systems, for uplift-based recommendation, the treatment is defined as recommendation, the outcome is defined as user behaviors, and the system aims to maximize the increment of user behavior caused by recommendation. However, the item cannot be both recommended and not recommended in the factual world, therefore, it is necessary to apply counterfactual into recommendation. Counterfactual has been widely applied into recommender systems to address practical issues and made great success. We will demonstrate details in this survey.

4 Causal Assumptions in Recommendation

In this section, we will introduce commonly used assumptions in causal inference [38].

Definition 4. ***(Stable Unit Treatment Value Assumption (SUTVA))*** *The potential outcomes for any individual do not vary with the treatment assigned to other individual, and, for each individual, there are no difference forms or versions of each treatment level, which lead to different potential outcomes.*

This assumption emphasizes the independence of each individual, which means that there are no interconnections between individuals. In recommendation, the individual usually represents the user. The traditional recommendation implicitly assumes the independence between users, which satisfies the SUTVA assumption. However, this assumption does not always hold in practical recommender systems. For example, in the recommendation for social networks, the users may connect with each other through the network structure. Some recommendation models do not have explicit users, for example, in session-based recommendation. In this case, the individual may be considered as the sessions, which temporally connect with each other.

Definition 5. ***(Ignorability)*** *Given the variables W, which are not affected by the treatment, treatment assignment X is independent to the potential outcomes, i.e., $Y(1), Y(0) \perp\!\!\!\perp X|W$.*

The ignorability assumption is also named as the unconfoundedness assumption. This assumption defines the treatment assignment under certain condition. Specifically, for individuals with the same variables W, the treatment assignment is random. This assumption is accepted by many recommendation algorithms, however, in real-world recommender systems, there may exists unobserved variables affect both the treatment and outcome, which has been studied by existing works [77, 78].

Definition 6. ***(Positivity)*** *For any value of variables W, which are not affected by the treatment, the treatment assignment is not deterministic:*

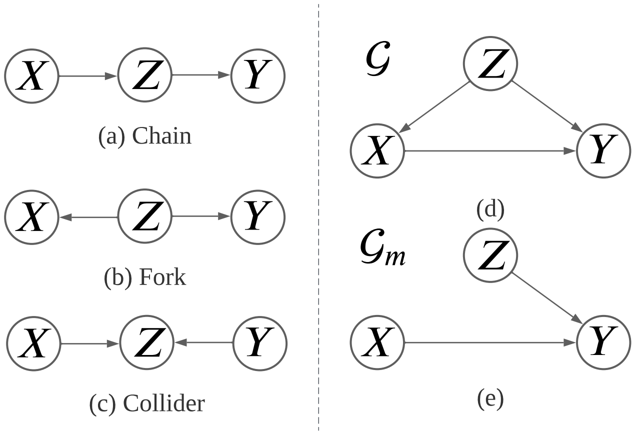
$$P(X = x|W = w) > 0, \quad \forall x \text{ and } w. \tag{1}$$

This assumption guarantees the feasibility and significance of estimating the treatment effect. If for some values of W, the treatment assignment is deterministic, then the outcomes of at least one treatment can not be observed forever for these values. In this case, estimating the treatment effect is impractical and meaningless. This assumptions hold in recommendation algorithm design. For each user, every items have the chance to be exposed to the users. Items that cannot be exposed are not within the research scope of recommender systems.

With above three assumptions, the connection between the observed outcomes and the potential outcomes can be established.

$$\begin{aligned} \mathbb{E}[Y(x)|W = w] =& \mathbb{E}[Y(x)|W = w, X = x] \\ =& \mathbb{E}[Y|W = w, X = x] \end{aligned} \tag{2}$$

Apart from above three commonly used assumptions, there is another way to represent the assumed mechanism.



(a) Chain

(b) Fork

(c) Collider

(d)

(e)

Fig. 1. X, Y, Z represent three variables. (a)-(c) show three fundamental causal graphs. (d) show and example of causal graph, and (e) represents the manipulated graph of (d) when intervene on variable X.

Definition 7. ***(Structural Causal Model (SCM))*** *A SCM consists of a set of endogenous (V) and a set of exogenous (U) variables connected by a set of functions (F) that determine the values of the variables in V based on the values of the variables in U.*

SCM is the key concept in the Pearl's causal framework, which provides stronger assumptions (than potential outcomes framework) about the mechanisms behind the scenarios, which indicates the relationships between variables other than the treatment and the outcome. Each SCM is associated with a graphical model $\mathcal{G}$, represented as a Directed Acyclic Graph (DAG), where each node is a variable in U or V and each edge is a function f. Each edge corresponds to a causal assumption: If the variable Y is the child of a variable X, then it is assumed that X is the direct cause of Y; If the variable Y is the descendant of a variable X, then it is assumed that X is the potential cause of Y. The causal graph is the key difference between potential outcomes framework and structural causal model framework, where potential outcomes framework does not consider the causal graph to depict causal relationships. However, we think that both frameworks are built on some assumptions, and the causal graph is just a stronger assumption, which cannot completely separate two frameworks. We introduce three fundamental causal graph in Figure 1.

Causal graph is a straightforward way to represents the underlying mechanism of recommender systems, and three typical causal graphs in Figure 1 often appear in the mechanisms of recommender systems. For example, the chain structure in Figure 1(a) appears in [77], where item decides intrinsic item features and intrinsic item features further decides user preference; the fork structure in Figure 1(b) appears in [79], where item popularity is considered as a common cause of both item exposure and interaction probability; the collider structure in Figure 1(c) appears in [80], where user click is the common outcome of both user interest and conformity. For SCM, an existing work [81] has shown that the traditional recommendation and causal recommendation can be unified through a causal view, where the recommendation models aim to estimate

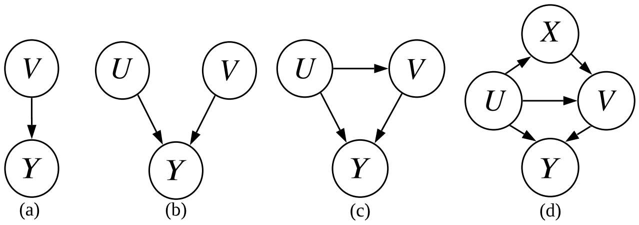


Fig. 2. Many traditional recommendation and causal recommendation can be unified under different causal graphs. In the graphs, U is user, V is item, X is user interaction history, Y is preference score. (a) Causal graph for non-personalized models. (b) Causal graph for similarity matching-based CF models. (c) Causal graph that considers the causality from user to item [82]. (d) Causal graph used in [81].

$P(Y|U, do(X))$ (i.e., Y represents the user preference, U denotes users, V denotes items) but with different causal graphs. More details can be found in Figure 2.

As we mentioned before, the *intervention* on the treatment variable can be interpreted as applying do-operation on the treatment variable. Intuitively, the do-operation means directly intervention, which cut off the influence from other variables to the treatment. Therefore, considering two variable X and Y, the desired interventional probability $P(y|do(x))$ can be intuitively calculated as $P_m(y|x)$, which is the observed probability on the manipulated graph. Specifically, the manipulated graph removes all the income edges to the treatment variable. For example, considering a simple causal graph as Figure 1 (d), where X is the treatment, Y is the outcome, and Z is the confounder, $P(y|do(x))$ on the original causal graph $\mathcal{G}$ is the same as $P_m(y|x)$ on the manipulated graph $\mathcal{G}_m$ shown as Figure 1 (e). An example in recommendation is taking intervention on item exposure, which generate the data of randomized experiments (i.e., data generation process follows the manipulated causal graph). We will introduce more details about randomized experiments in Section 6. Similar to the intervention on causal graphs, the intervention on structural equations take the intervened value as the input to calculate the output of the structural equations.

The introduced assumptions bridge the gap between the observed correlation and the estimated causation. We will introduce some commonly used methods based on introduced assumptions.

5 Causal Effects in Recommendation

After introducing the basic representation of the causal representation, many different kinds of causal effects can be defined using basic representations.

A basic causal effect is called as the treatment effect (i.e., the outcome change if another treatment has been applied). More specifically, the treatment effect can be measured at the population, treated group, subgroup and individual levels. Here we define the treatment effect under binary treatment to make it clear, and it can be extended to multiple treatments by comparing their potential outcomes [38]. We takes the potential outcome as an example, and the do-operation can be applied in similar ways.

The treatment effect at the population level is named as the **Average Treatment Effect (ATE)** (some reference also name it as the **Average Causal Effect** [68] or the **Total Effect** [83]), which is defined as:

$$ATE = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \tag{3}$$

The treatment effect at the treated group level is called **Average Treatment effect on the Treated Group (ATT)** (some reference also name it as **Effect of Treatment on the Treated (ETT)**[27, 68]), which is defined as:

$$ATT = \mathbb{E}[Y(1)|X = 1] - \mathbb{E}[Y(0)|X = 1] \tag{4}$$

where $Y(1)|X = 1$ and $Y(0)|X = 1$ represent the potential outcomes under both treatments of the treated group.

For the subgroup level, the treatment effect is named as **Conditional Average Treatment Effect (CATE)**, which is defined as:

$$CATE = \mathbb{E}[Y(1)|W = w] - \mathbb{E}[Y(0)|W = w] \tag{5}$$

where W denotes the variables (i.e., grouping by multiple variables) defining the subgroup which are not affected by the treatment, and $Y(1)|W = w$ and $Y(0)|W = w$ are the potential outcomes under both treatments within the subgroup with $W = w$.

At the individual level, the treatment effect is called as **Individual Treatment Effect (ITE)**, which can be represented as:

$$ITE = Y_i(1) - Y_i(0) \tag{6}$$

where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes for treatment $X = 1$ and $X = 0$ of individual i respectively. The ITE is considered equivalent as the CATE [84, 85] if each subgroup represents an individual.

The treatment effect on different level has been used as quantitative evaluation in recommender systems to handle many issues. For example, ITE is used to estimate the uplift value of recommendation [75, 86, 87, 88]; ATE can be used to evaluate explanations [89] and estimate unbiased preference [90]; ATT is used to evaluate counterfactual fairness [91]; etc.

In addition to the treatment effect at different levels we introduced above, there are some causal effects for mediation analysis. A mediation model seeks to explain the process that underlines a causal relationship between the treatment and the outcome via the inclusion of a third variable, known as a mediator variable. Let X, Y, and M denote treatment, outcome, and mediator respectively. We will introduce three types of effects under the binary treatment for mediation analysis.

First, Controlled Direct Effect (CDE) measure the expected increase in Y as the treatment changes, while the mediator is set to a specific value m for the entire population, which can be defined as:

$$CDE(m) = \mathbb{E}[Y|do(X = 1, M = m)] - \mathbb{E}[Y|do(X = 0, M = m)] \tag{7}$$

Second, Natural Direct Effect (NDE) measures the expected increase in the outcome as the treatment changes, while the mediator is set to whatever value it would have attained prior to the change, i.e., $X = 0$, which can be defined as:

$$NDE = \mathbb{E}[Y|do(X=1, M=M_0)] - \mathbb{E}[Y|do(X=0, M=M_0)] \tag{8}$$

where M_0 represents the value of mediator under treatment as 0.

Finally, Natural Indirect Effect (NIE) measures the expected increase in outcome when the treatment is held constant at $X = 0$, while M changes to whatever value it would have attained under $X = 1$, which can be defined as:

$$NIE = \mathbb{E}[Y|do(X=0, M=M_1)] - \mathbb{E}[Y|do(X=0, M=M_0)] \tag{9}$$

where M_1 represents the value of mediator under treatment as 1. NIE captures the portion of the effect which can be explained by mediation alone.

The above direct and indirect effects play an important role in recommendation as well. The direct and indirect effects help the models quantitatively evaluate path-specific effects to detect and remove undesired effects. For example, they can be used to identify direct and indirect discrimination to achieve or explain fairness [92, 93], they can be used to identify and remove some bias [94, 95], etc.

6 Causal Estimation Methods in Recommendation

Having defined the causal effects, the next logical step is to ask how can we estimate those effects. One way is to perform a randomized experiment.

6.1 Randomized Experiments.

To measure the average treatment effect, an ideal way is to apply different treatment to the same group of individuals. However, the ideal solution is impractical in real-world situation. It can only be approximate by a randomized experiment. Specifically, a randomized experiment randomly assigns individuals into the treated group or the control group. The estimated ATE can be obtained by the difference of the average outcomes of two groups. To understand why a randomized experiment is the golden standard for estimating the average treatment effect, it is necessary to understand how correlation is different from causation.

$$\begin{aligned}
&\mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0] \\
\overset{1}{=}&\mathbb{E}[Y(1)|X=1] - \mathbb{E}[Y(0)|X=0] \\
\overset{2}{=}&\underbrace{\mathbb{E}[Y(1)|X=1] - \mathbb{E}[Y(0)|X=1]}_{\text{ATT}} \\
&+\underbrace{\mathbb{E}[Y(0)|X=1] - \mathbb{E}[Y(0)|X=0]}_{\text{bias}}
\end{aligned} \tag{10}$$

Here, step 1 follows the fact that $Y(1)$ is the observed outcome when conditioning on $X = 1$ and $Y(0)$ is the observed outcome when conditioning on $X = 0$; step 2 adds and subtracts $\mathbb{E}[Y(0)|X = 1]$ to construct the ATT term and the bias term. The bias term in Eq.(10) creates the gap between the correlation and causation. The randomized experiment eliminate the bias term by randomly assigning individuals into the treated group or the control group. More specifically, the random assignment makes the potential outcomes are independent from the treatment $Y(1), Y(0) \perp\!\!\!\perp X$ (it does not imply that outcomes are independent from the treatment), thus $\mathbb{E}[Y(0)|X = 1] = \mathbb{E}[Y(0)|X = 0]$. Given $Y(1), Y(0) \perp\!\!\!\perp X$, Eq.(10) can be rewrite as:

$$\mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \tag{11}$$

Therefore, a randomized experiment can simply estimate ATE as the difference of the average outcomes of the treated group and the control group. In recommendation, the randomized experiments are usually used to handle the bias [82, 96, 97, 98, 99, 100, 101, 102]. Specifically, by taking item exposure as the treatment, the randomized experiments follow the random recommendation policy instead of the deployed policy, in which return the unbiased data (i.e., also called as uniform data) for recommendation.

A randomized experiment is not a one-size-fits-all solution for causal inference. In reality, randomized experiments are always time-consuming and expensive, thus the study usually involve small number of individuals, which may not be representative of the population. Meanwhile, ethical concerns largely limit the applications of the randomized experiments such as environmental health studies. In addition, the randomized experiments cannot explain the causation on the individual level. Therefore, given the wide availability of observational data, the observational study is a shortcut for causal inference.

6.2 Observational Data

Although the observational study could be a shortcut for causal inference, there are some issues of the observational data should be carefully considered during designing the causal models. The existence of confounders is a critical problem in the observational data.

Definition 8. ***(Confounders)*** *Confounders are variables that affect both the treatment assignment and the outcome.*

Due to the existence of confounders, some spurious effect may be observed (taking relationship between ice cream consumption and shark attacks as an example). Confounders widely exists in recommender systems. The existence of confounders often results in different bias based on the definition of confounders. For example, taking item popularity as a confounder, it will lead to popularity bias [79]. In addition to some observed and measurable confounders, such as item popularity, some unobserved or immeasurable confounders (i.e., which violate the ignorability assumption in Section 4) exist in real-world recommendation and have been widely studied by the community [74, 78, 81].

Simpson's paradox is another phenomenon that could be observed in the observational data. From Table 1, it can be observed that in both male and female groups, taking the drug has a better recovery rate; but in the total population, not taking drug has a better recovery rate. This phenomenon is usually caused by confounders. The Simpson's paradox can be also observed in recommender systems [103].

TABLE 1
Results of a study into a new drug, with gender being taken into account [68]. a/b represents a out of b recovered.

	Drug	No Drug
Male	81/87 (93%)	234/270 (87%)
Female	192/263 (73%)	55/80 (69%)
Total	273/350 (78%)	289/350 (83%)

Macdonald [103] observes the Simpson's paradox in offline evaluation for recommendation, and propose a method to mitigate the paradox in offline evaluation.

Compared to the experimental data, observational data only provides the information about what has occurred, but the why a specific treatment is token is unknown. Given that the treatment assignment mechanism is unknown, the bias term in Eq.(10) cannot be eliminated or quantitatively measured. Therefore, the bias caused by unknown treatment assignment is also a critical issue that should be carefully handled in model design.

6.3 Methods Relying on Assumptions

In some complex scenarios, it is risky to assume the causal mechanism based on prior knowledge. In this case, SUTVA, ignorability and positivity assumptions support some methods to estimate the potential outcomes.

One commonly used method is based on the idea of reweighting. As we mentioned before, due to the unknown treatment assignment mechanism, there may exists the bias problem. By assigning appropriate weight to each sample in the observational data, a pseudo-population can be created on which the distributions of the treated group and the control group are similar. There are two commonly used reweighting methods: inverse propensity scoring and confounder balancing.

Definition 9. *(Propensity Score) The propensity score is defined as the conditional probability of treatment given background variables:*

$$e(w) = P(X = 1|W = w) \tag{12}$$

Given the propensity scores defined above, inverse propensity scoring methods [104, 105] assign a weight based on propensity score to each observed samples. Thus the estimated ATE based on the observed samples can be rewrite as:

$$ATE_{IPS} = \frac{1}{n_1} \sum_{i,x_i=1} \frac{y_i}{e(w_i)} - \frac{1}{n_0} \sum_{j,x_j=0} \frac{y_j}{1-e(w_j)} \tag{13}$$

Inverse propensity scoring (IPS) is often used to design unbiased estimator for recommender systems [90, 106], where the propensity score can be pre-defined or learned from the data.

Although the use of propensity score is effective to reduce the bias, there are some issues during applying IPS in practice. First, the correctness of the IPS estimator highly relies on the correctness of the propensity score estimator. To handle this dilemma, some augmented IPS methods are proposed, such as doubly robust estimator [107]. Another drawback is that the IPS estimator has variance problem, that the estimator is unstable if the estimated propensity scores are small. To overcome this drawback, some methods propose to clip the propensity score [70] or trim samples with small propensity scores [108].

Another reweighting method is confounder balancing [109, 110, 111]. The motivation is that the confounders can be balanced by the moments, which uniquely determine the distribution of variables. Thus the sample weights can be learned to estimate the causal effect through reweighting. The confounder balancing based methods are used for stable learning [112] and robust recommendation [113].

In addition to reweighting methods, stratification is another representative method. The idea of stratification is to split the entire population into homogeneous subgroups, which makes the treated group and the control group are similar in each subgroup. Ideally, in this case, the samples in the same subgroup can be viewed as sampled from the data under randomized experiments. Macdonald [103] adopt this idea to mitigate Simpson's paradox in offline evaluation for recommendation.

In some applications, the causal mechanism is safely assumed based on prior knowledge or expert knowledge. In this case, the causal mechanism can be represented as a SCM as we introduced before. Although structural causal model framework requires stronger assumptions than potential outcomes framework, it also enable reasoning through the graph. Using a SCM, the key difference between causation and correlation is *do*-operations, which is the basic element to estimate the causal effect. As we mentioned, the *do*-operation can estimated by manipulated graph. However, the data from the manipulated graph is generated from randomized experiments. The approaches based on the data generated by the original causal graph are useful in practice. Applying backdoor adjustment is a popular approach.

Definition 10. ***(Back-door Criterion)*** *A set of variables Z satisfies the backdoor criterion related to an ordered pair of variables (X, Y) in a causal graph $\mathcal{G}$ if Z satisfies both (1) No node in Z is a descendant of X and (2) Z blocks every path between X and Y that contains an arrow into X.*

Through identifying a set of variables satisfying the back-door criterion, the causal effect can be estimated using back-door adjustment formula.

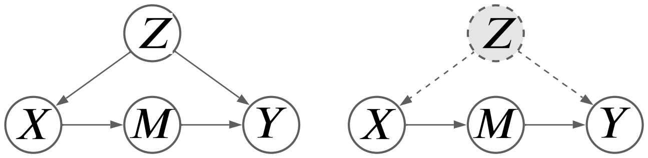
Definition 11. ***(Back-door Adjustment)*** *If a set of variables Z satisfy the back-door criterion related to an ordered pair of variables (X, Y), and if $P(x, z) > 0$, then the causal effect of X on Y is identifiable and is given by*

$$P(y|do(x)) = \sum_z P(y|x, z)P(z) \tag{14}$$

Given the population of the observed data, if we divide the subgroup based on value of Z, Eq.(14) can be considered as calculating the causal effect by the weighted sum of each subgroup, which is very similar to the stratification methods. Additionally, Eq.(14) can be rewrite as:

$$P(y|do(x)) = \sum_z \frac{P(y, x, z)}{P(x|z)} \tag{15}$$

where $P(x|z)$ is known as the "propensity score", therefore, the back-door adjustment is also an alternative representation of IPS methods. The back-door adjustment is widely



(a) (b)

Fig. 3. (a) An example of applying back-door adjustment on the causal graph. (b) An example of causal graph with an unobserved confounder, in which the causal values can be estimated by the front-door adjustment.

used to address issues in recommendation, such as bias issues [79, 114], echo chambers [115], etc.

When we consider the do-operations, the interventions are not limited to actions that force a variable or a group of variables to take on specific value. In general, interventions may involve dynamic policies in which the treatment variable X is made to respond in a specified way to some set Z of other variables, which is denoted as $x = g(z)$. In this case, the estimated causal effect $P(Y = y|do(X = g(Z))$ can be calculated as:

$$
\begin{aligned}
&P(Y = y|do(X = g(Z)))\\
=&\sum_z P(Y = y|do(X = g(Z)), Z = z)P(Z = z|do(X = g(Z)))\\
=&\sum_z P(Y = y|do(X = g(Z)), Z = z)P(Z = z)\\
=&\sum_z P(Y = y|do(X = x), Z = z)|_{x=g(z)}P(Z = z)
\end{aligned}
\tag{16}
$$

In recommendation, the feedback data is collected from a deployed recommendation algorithm, thus the recommendation policy exists in the data generation process. Considering the dynamic policy as the recommendation policy, conditional intervention can also be applied to design causal recommendation models [81]. In recommendation scenario, observing an interaction in the feedback data does not imply that the interaction is destined to happen, thus the causal adjustment methods is sometimes applied with counterfactual reasoning [77, 81, 115].

Apart from above adjustment formulas, there are some rules are valid for interventional probabilities, which are called as the rules of do-calculus. Before introducing the specific rules, we first introduce some notations. Let X, Y, Z, and W be arbitrary disjoint sets of nodes in a causal DAG $\mathcal{G}$. $\mathcal{G}_{\overline{X}}$ denotes the graph obtained by deleting from $\mathcal{G}$ all arrows pointing to nodes in X. Likewise, $\mathcal{G}_{\underline{X}}$ denotes as the graph obtained by deleting from $\mathcal{G}$ all arrows emerging from nodes in X. The rules of do-calculus can be represented using above notations.

Definition 12. ***(The rules of*** ***do*-*calculus)*** *The following three rules are valid for every interventional distribution compatible with a causal graph* $\mathcal{G}$

Rule 1 *(Insertion/deletion of observations):*

$$
\begin{aligned}
&P(y|do(x), z, w) = P(y|do(x), w)\\
&\textit{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{\mathcal{G}_{\overline{X}}}
\end{aligned}
\tag{17}
$$

Rule 2 *(Action/observation exchange):*

$$
\begin{aligned}
&P(y|do(x), do(z), w) = P(y|do(x), z, w)\\
&\textit{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{\mathcal{G}_{\overline{X}\underline{Z}}}
\end{aligned}
\tag{18}
$$

Rule 3 *(Insertion/deletion of actions):*

$$
\begin{aligned}
&P(y|do(x), do(z), w) = P(y|do(x), w)\\
&\textit{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{\mathcal{G}_{\overline{X}\,\overline{Z(W)}}}
\end{aligned}
\tag{19}
$$

where $Z(W)$ *is the set of* Z*-nodes that are not ancestors of any* W*-nodes in* $\mathcal{G}_{\overline{X}}$*.*

With the help of the rules of do-calculus and introduced adjustment formulas, the interventional probabilities can be estimated by the observational data.

6.4 Methods with Relaxed Assumptions

Although above methods relying on introduced assumptions basically satisfy the requirement of estimating causal effect from the observational data, in practice, for some specific applications, the introduced assumptions may not always hold. There are some methods trying to estimate the causal effect with relaxed assumptions.

SUTVA assumes that individuals are independent and identical distributed. However, in some real-world applications, such as social networks, SUTVA cannot hold anymore since individuals are inherently interconnected with each other through the network structure. To handle this issue in real applications, a commonly used approach is applying a model, which capture the interconnection, into a causal inference model. For examples, applying graph convolutional networks into a causal inference model to handle the network data [116].

The ignorability assumption assumes that the treament assignment is independent to the potential outcomes given the background variables. However, it is impossible to identify and collect all the background variables in real world, thus the ignorability assumption is hard to satisfy. In other words, there may exist unobserved confounders as we mentioned before. Only using observational data to estimate the causal effect is difficult, an alternative way is to combine the limited experimental data and observational data together [117]. In recommendation, the unbiased data is collected from randomized experiments, using a small part of unbaised data and a large part of observed feedback is a popular way to design unbiased recommendation models.

Another solution is based on the assumed SCM, which models the unobserved confounders into the causal graph (an example is shown in Figure 3(c)). Similar to applying the back-door adjustment, we first identify a set of variables satisfying the front-door criterion.

Definition 13. ***(Front-door Criterion)*** *Given an ordered pair of variables* (X, Y) *in a causal graph* $\mathcal{G}$*, a set of variables* Z *satisfies the front-door criterion with respect to* (X, Y) *if* Z *satisfies the following conditions:*

- *Z intercepts all directed paths from X to Y.*
- *There is no unblocked back-door path from X to Z.*
- *X blocks all back-door paths from Z to Y.*

Given a set of variables that satisfies the front-door criterion, we can identify the causal effect with unobserved confounders [68].

Definition 14. ***(Front-door Adjustment)*** *If a set of variables Z satisfy the front-door criterion related to an ordered pair of variables (X, Y), and if $P(x, z) > 0$, then the causal effect of X on Y is identifiable and is given by*

$$P(y|do(x)) = \sum_{z} P(z|x) \sum_{x'} P(y|x', z)P(x') \tag{20}$$

The existence of unobserved confounders is widely recognized by the community [74, 77, 78, 118], there are some works [77, 118] that attempt to apply front-door adjustment in recommendation.

Using instrumental variables is a possible way to get around the ignorability assumption and conduct causal inference. Instrumental variables are defined as variables that only affect the outcome via the treatment variables. Typical instrumental variables methods [119, 120] adopt two-stage models: the first stage reconstructs the treatment variable based on the instrumental variable and the second stage reconstructs the outcome based on the treatment from the first stage. In recommender systems, Si et al. [121] adopt the instrumental variable to design a model-agnostic recommendation framework using search data.

7 Causal Discovery in Recommendation

The above methods aim to learn the causal effect, there is another branch of causal models targeting at learning causal relations, which is also known as causal discovery. Except for few works only aim to identify treatment and outcome [122], most of the works aim to discover causal graphs. Following [39, 123], traditional methods can be divided into three categories: constraint-based, socre-based and those based on functional causal models.

Constraint-based Algorithms learn a set of causal graphs that satisfy the conditional independence embedded in the data and statistical tests are utilized to verify if a candidate graph satisfies the independence. Score-based algorithms learn causal graphs by maximizing the scoring function $S(\mathbf{X}, \mathcal{G})$, which returns the score of the causal graph $\mathcal{G}$ given data $\mathbf{X}$. Algorithms based on Functional causal models (FCMs) usually define a variable as a function of its directed causes and some noise term (e.g., linearly weighted by the adjacency matrix of the causal graph [124]) and optimize the designed objective to learn the parameters of the functions. We only briefly introduce the causal discovery methods, interested readers may refer to [39, 123] for more details.

Most existing works in causal recommendation are based on pre-defined causal graph representing the underlying causal mechanisms. The pre-defined causal graphs are usually defined based on expert knowledge, which may be inaccurate and quite simple (i.e., only involve few variables). Leveraging causal discovery in recommendation will handle these issues. There exist few works [125, 126] design recommender systems with causal discovery techniques based on continuous optimization [127]. The learned causal mechanism will increase the explainability of recommender systems and guide the model design for other aspects, such as fairness, unbiasedness.

8 Causal Explainability in Recommendation

With the development of machine learning, accuracy is no longer the only only pursuit. Moreover, transparency and trustworthiness start to obtain increasing attention. For example, heathcare AI is required to provide not only accurate diagnoses, but also supporting explanations to convince patients. Recommender systems, with humans in the loop, also require transparency. Explainable recommendation, which emerged and developed with the pursuit of transparency and trustworthiness of recommender systems, has been increasing popular in both academia and industry. It aims to provide explanations for the recommended items, which will benefit the community in many ways. For consumers, the explainable recommendation is able to help them make better decisions. For the platform, it may improve the transparency, persuasiveness, trustworthiness and user′s satisfaction of the system. For model developers, it is an important tool to understand the designed model and accelerate the design cycle. In this section, we will first introduce the overview of the explainable recommendations, and then summarize the existing causal methods, as well as some open problems related to causal inference.

8.1 Problem Introduction

The research of explainable recommendation, as a sub-area of explainable AI, was proposed and defined by [128]. With the rapid development of deep neural networks, the state-of-the art recommender systems widely adapt deep models to improve the recommendation performance. However, these deep models are too complicated for users to understand the decision made by the intelligent systems, thus a deep model is usually considered as a black-box. Recommender systems, serve as essential decision-making systems in daily life, are required to provide accurate decision results as well as underlying reasons. For example, a stock investor needs to know which characteristics lead to the recommendation before making the final decision. A consumer hopes to understand why the recommended items are worth buying before paying.

The explainable models can be either model-intrinsic or model-agnostic. The former one refers to generating explanations simultaneously with the recommendation results and the later one refers to generating explanations after providing the recommendation results. Model-intrinsic (also known as ad-hoc) explainable models usually design the explanation generation mechanism as a part of decision-making process, and model-agnostic explanations (also known as post-hoc) explainable models usually design separate mechanisms for generating explanations.

The explanations can be presented in many different ways, which usually depend on what kind of information source is used for explanations. Typically, the explanations can be presented as related users or items [89, 129], the features of users or items [128, 130], generated textual sentence [131, 132], visual explanations [133], graph [134], etc. Existing works have made many successes in explainable recommendations with different information sources. For example, Zhang et al. [128] propose Explicit Factor Model (EFM) which extract explicit item features and user opinions from user reviews to provide feature-level explanations. Peake and Wang [129] extract association rules to provide purchased items as an explanation in a model-agnostic manner. Xian et al. [134] perform explicit reasoning path with knowledge graph to provide recommendations and explanations. In addition, Existing works have introduced explainability into conversational recommendation. Chen et al. [135] develop an Explainable Conversational Recommendation (ECR) model to provide accurate recommendations as well as high quality explanations by multi-round conversations. Incorporating causal inference ideas and techniques brings new opportunities for explainable recommendations. In the following part, we will focus on causal-related methods. Interested readers may refer other surveys [29, 30] for more explainable recommendation approaches.

8.2 Causal Methods

8.2.1 Counterfactual

In recent years, counterfactual reasoning draws more and more attention in explainable AI. For any AI system that makes predictions based on machine learning models, no matter white-box or black-box, counterfactual reasoning looks for what input (e.g., aspects, features) should be changed, and by how much, to acquire a different prediction. Then, the altered input will comprise the explanation. For instance, when generating explanations for a rejected loan request, it could be something like: if your annual income is $50,000$, instead of $30,000$, your request will not be rejected. Some existing works have introduced the idea of counterfactual reasoning into recommendation scenarios for generating explanations, which looks for minimal changes in the recommender system (e.g. item features, items in the history, user's behaviors, etc.) leading a different prediction to identify the most essential part (e.g. item features, items in the history, user's behaviors, etc.) as the explanations. Ghazimatin et al. [136] generate explanations for a recommender system based on users' actions in in the history. More specifically, it introduces a searching algorithm on a knowledge graph to look for the minimal set of user's history to be cut off, such that the user will receive different recommendation results. Tan et al. [130] proposes a counterfactual explanation framework for generating feature-level explanations. It introduces two new concepts, explanation complexity and explanation strength. These two concepts are used to formulate a counterfactual optimization problem, as well as an evaluation metric to evaluate the generated explanations. Later in [137], a similar counterfactual explanation framework is also used to explain which features are causing fairness issues in recommender system. Tran et al. [138] utilizes an influence function to analyze the training data. Then, a counterfactual set of training data are used for generating explanations.

8.2.2 Causal Discovery

Explainable recommendation models based on causal discovery are still in theirs infancy. Causal discovery methods aim to extract causal relations among variables from the data. Existing causal discovery based approaches in recommendation provide model-intrinsic explanations. More specifically, through the extracted causal relations, the causal discovery based recommendation models are able to provide recommendations simultaneously with corresponding causal relations as explanations. As we mentioned, causal discovery methods usually try to learn a causal graph. In recommendation scenario, considering the extremely large amount of items, the learned causal graphs are typically based on item group level. For example, Wang et al. [125] propose to learn a cluster-level causal graph to guide sequential recommendation. Based on the learned cluster-level causal graph and cluster assignment for each item, the model is able to calculate the causal relations between items. The item in interaction history with the strongest causal relation with the recommended item is identified as the explanation. Xu et al. [126] aim to learn a causal graph on product type (PT) level for PT-level recommendation. Particularly, the model takes collected feedback data as the result of the mixture of two completing mechanisms: a causal mechanism based on user intention and a intervention mechanism based on deployed recommendation algorithm. The recommendation and corresponding explanations are generated via the learned PT-level causal graph.

8.3 Open Problems

Despite the above successful usages in causal explainable recommendation, there are open problems that expected to be solved in the future. First, causal discovery based explainable recommendation models, are capable of generating model-intrinsic explanations, need further exploration. Second, the current counterfactual explanation algorithms are claimed to be model-agnostic because they are able to be applied on any recommendation models (or at least a wide range of recommendation models). However, the model itself has to be reachable. It is not certain about how to apply counterfactual explanation algorithms on an recommendation model that are not accessible by the algorithm user. Finally, there are currently no methods to leverage other causal reasoning methods, such as the do-calculus, to generate explanations.

9 Causal Fairness in Recommendation

Recommender system, as a powerful tool for business, has been widely used to improve user engagement and further create higher profit. Classical recommender systems mainly care about how to precisely estimate user preferences. However, in recent years, concerns about fairness in recommendation have attracted much attention from both industry and academia [31, 32, 139, 140, 141]. With the development of recommendation techniques, recommender systems have been widely used to assist or even replace human decision-making in several domains. Several studies have shown that the unfairness may lead to negative consequences [142, 143, 144], which in turn may have significant social impacts. For example, in e-commerce, the unfairness of exposure of items may hurt the benefits of the platform and providers in long-term [145]; in educational recommendation [146], an unfair system due to gender imbalance [147] may discourage females from selecting STEM (i.e., science, technology, engineering, and mathematics) topics, which may affect society for generations; An unfair ad recommendation may even result in racial discrimination [148]. Therefore, to increase the applications of recommender systems and maintain a healthy social impact, it is critical to consider fairness in recommendations and build a reliable decision-making system.

9.1 Problem Introduction

Before achieving fairness in recommender systems, one should first understand the reasons of unfairness. Bias and discrimination are two commonly accepted causes of unfairness [31, 32, 33, 149]. Biases in recommender systems mainly consist of bias in data and bias in algorithm. The bias in data may come from data generation, collection, sampling, and storage. For example, in recommender systems, the training data is collected from a deployed system, if the algorithm underlying the deployed system makes biased predictions, then the generated data may involve biases. The bias in data may affect the algorithms, since most machine learning algorithms rely on data to be trained and make predictions after training. If the training data contains biases, the algorithms trained on them will learn biased knowledge from these biases and further lead to unfairness. For example, if the training data shows significant imbalance between majority user/item group and minority user/item group, it is high likely that the recommendation algorithm learns much better on the majority group and results in discrimination on the minority group. Except for the bias in data, the recommendation algorithm itself may enhance existing biases and cause unfairness, which is referred to the bias in algorithm. For example, some recommendation algorithms may enhance the popularity bias, where popular items will get more recommendation than less popular items with equal or similar quality. Discrimination, as a multidisciplinary problem [150, 151, 152], is also a cause of unfairness defined as an unjustified difference in treatment on the basis of any physical or cultural characteristic (e.g., race, gender, etc.) due to human prejudice and stereotyping. It is worth mentioning that unfairness is not only caused by bias and discrimination. For example, there may exist conflicts or trade-offs between different kinds of fairness [31, 33, 153], where achieving one fairness will hurt another fairness.

To fight against unfairness, it is important to define fairness. In general machine learning, fairness can be defined on target level (i.e., to achieve fairness on group-level or individual-level). Specifically, fairness can be categorized into group fairness and individual fairness.

- **Group Fairness**: Group fairness defines the fairness on group-level, which is based on the idea that different groups should be treated equally. Here the groups can be divided in many ways, where the most commonly used way is to split the groups based on some explicit sensitive attributes.
- **Individual Fairness**: Individual fairness defines fairness on individual-level, which is based on the idea that similar individuals should receive similar predictions. Moreover, individual fairness can be theoretically considered as a very special group fairness, which divides each individual into different groups.

Since fairness in recommender systems relates to the benefits from multiple stakeholders [144, 154, 155, 156, 157], the request of fairness may come from different sides. Therefore, the definition of fairness in recommendation can also be divided into user-side fairness and item-side fairness.

- **User-side Fairness**: User-side fairness aims to satisfy the fairness requirements from users (consumers). The request from the user side are mainly focusing on the recommendation quality (i.e., recommendation performance). The user-side fairness can be achieved on both group-level and individual-level. User-side fairness on group-level aims to reduce the discrepancy of recommendation quality between different user groups, where the user groups are divided by sensitive features, such as race or gender [142, 158], or by assigned features (e.g., cold users vs. heavy users [159], active users vs. inactive users [140, 160]). For user-side fairness on individual-level, the recommendation quality should be unchanged even an individual's sensitive features have changed. For example, Li et al. [139] incorporate the idea of counterfactual fairness [91] to design a recommendation model which makes the recommendation performance unchanged even the user's sensitive features are flipped in the counterfactual world.
- **Item-side Fairness**: Item-side fairness aims to satisfy the fairness from items side, which mainly focuses on requesting equal exposure opportunity of items to maintain market fairness. Here the items refer to "items" to be ranked or recommended. For example, in e-commerce, the items refer to products to be sold; in recruitment system, the items refer to job seekers (item providers). One branch of existing work focuses on achieving fairness according to item attributes. For example, some works [145, 161, 162, 163, 164, 165] achieve the fair exposure between popular and unpopular items to prevent unpopular items from being under-exposed. Moreover, another branch of research work mainly focuses on achieving fairness based on the sensitive attributes of item providers, such as gender [166, 167, 168], geographic provenience [169, 170, 171], etc.

It is worth noting that the user-side fairness and item-side fairness may not exclusive to each other, where two-sided fairness [172, 173, 174, 175, 176] approaches are proposed to satisfy the fairness demands from both sides. Besides the taxonomies mentioned above, there are also some taxonomies [31, 33] that are used to classify fairness in recommendation from other perspectives. For example, static fairness vs. dynamic fairness [143, 143, 177, 178]; short-term fairness vs. long-term fairness [145, 179]; populational fairness vs. personalized fairness [139, 180, 181]; blackbox fairness vs. explainable fairness [137], centralized fairness vs. decentralized fairness [182, 183].

Typically, the proposed approaches to achieve fairness in recommendations can be roughly divided into three categories: pre-processing methods, in-processing methods and post-processing methods [31, 33, 149, 184]. Pre-processing methods usually aim to achieve fairness by minimizing the bias in the data before the model training. Compared with other types of methods, there are fewer works on pre-processing methods. Some representative methods include fairness-aware data sampling approach to cover items of all groups, data balancing approach [185] to increase the coverage of minority groups and data repairing approaches to ensure label correctness and remove disparate impact [186]. In-processing methods propose to incorporate fairness requirements as a part of the objective function to achieve fairness during the training. Typically, the fairness requirement works as a regularizer or a constraint [66, 140, 145, 158, 162, 187, 188, 189, 190, 191]. To minimize the unfairness while minimizing the original loss function (i.e., recommendation accuracy loss), it is also important to find a trade-off between recommendation accuracy and fairness [145, 192], which is also sometimes formulated as a multi-objective learning problem [192]. Post-processing methods aim to achieve fairness in inference stage after the training, by techniques such as re-ranking [140, 193, 194, 195] or multi-armed bandit [196, 197, 198]. To measure the unfairness, many different fairness metrics are proposed. For example, Absolute Difference (AD) [66] measures the absolute difference between the performance of protected group and unprotected group; Normalized Discounted KL-divergence [199] calculates a normalized discounted cumulative value of KL-divergence for each position, etc. More possible fairness metrics can be found in [32].

Recently, researchers have noticed that fairness cannot be well detected by solely correlation or association. Specifically, fairness criteria are based on solely joint distribution of random variables [200], such as outcomes, features, sensitive attributes, etc. However, recent work [201] shows that any definition of fairness that depends merely on the joint probability distribution is not necessarily capable of detecting discrimination. Therefore, many approaches [91, 93, 200, 202, 203] are proposed to address the problem of unfairness through the lens of causality.

In general machine learning, causal-based fairness notations are mostly defined on intervention or counterfactual. To measure the unfairness in causal-based fairness, one challenge is understanding the causal relationships that account for different outcomes. Causal graph, as a powerful tool for causal reasoning, is usually used to represent the causal relationships among variables. Given the causal graph capturing the causal relationships, many causal effects are used to measure the unfairness. For example, ATE (as Eq.(3), also known as Total Effect [27]) is used to measure the effect of changing sensitive attributes to the outcomes, Kilbertus et al. [204] measure the indirect causal effects [205] from sensitive attributes to outcomes and eliminate the directed path from sensitive attributes to outcomes except via a resolving variable, where resolving variables refer to any variables in the causal graph that are influenced by sensitive attributes in a non-discriminatory way. More details of causal-based fairness notations can be found in [76]. Counterfactual fairness is a commonly used definition of fairness in causal-based fairness. Counterfactual fairness is an individual-level causal-based fairness notion, which requires that the predicted outcome should be the same in the counterfactual world as in the real world for any individual [91]. The basic idea is minimizing the ATT (as Eq.(4), some references also name it as ETT [27, 68, 132]) conditioned on all features to receive the same probability distribution in the factual and counterfactual world. For counterfactual fairness in recommendation, the definition is given as follows [139]:

Definition 15. *(Counterfactual Fairness in Recommendation) The counterfactual fairness is satisfied for a recommendation model if for any user u with sensitive attributes $Z = z$ and remaining features $X = x$:*

$$P(L_z|X = x, Z = z) = P(L_{z'}|X = x, Z = z) \tag{21}$$

for all L and any value z' attainable by Z, where L denotes the top-k recommendation list for user u.

In the next section, we will introduce some causal methods for fairness in recommendation.

9.2 Causal Methods

9.2.1 Reweighting

As we mentioned before, bias is a widely accepted cause of unfairness, thus some existing work adopts inverse propensity scoring (IPS) methods to solve the bias in recommendation. For example, popularity bias will lead to item-side unfairness that popular items may obtain more exposure opportunities. The IPS approaches the biases are caused by non-randomly assigned treatment, thus use the inverse propensity to reweight the samples to remove the bias. For example, Schnabel et al. [90] consider recommendation as treatment and apply an IPS estimator in an Empirical Risk Minimization framework for learning to solve bias in recommendation. Saito et al. [70] design an IPS-based estimator for unbiased pairwise learning. Wang et al. [106] use a small part of unbiased data to train a propensity model and use biased data to train an IPS-based rating model. The IPS-based approaches are easy to implement but it requires an accurate propensity estimator and suffers from high variance [206, 207].

Although biases in data are commonly recognized as the main causes of unfairness in recommendation, the relationship between bias and fairness has not been clearly understood or discussed. More specifically, the debiasing methods are usually proposed to improve the recommendation performance by removing bias, thus the models are evaluated by recommendation metrics instead of fairness metrics. Many works on fairness are not implemented by debiasing methods but directly designed by fairness requirements, which may result in a trade-off between accuracy and fairness. In the following part, we focus on fairness methods. More discussion of debiasing methods can be found in Section 12.

9.2.2 Counterfactual

Counterfactual fairness, as a causality-based definition of fairness, requires the predicted outcomes to be the same in the counterfactual world as in the factual world. To achieve counterfactual fairness, it is important but challenging for a fair model to predict the outcomes in the counterfactual world (i.e., the sensitive attributes have been changed). Ma et al. [208] propose a counterfactual data augmentation module, which is trained based on a variational auto-encoder with a fairness constraint, to generate counterfactual data with different sensitive attributes. By maximizing the similarity between the representation learned from the original data and the different counterfactual data, the designed model is able to achieve counterfactual fairness. Mehrotra et al. [209] use counterfactual estimation to evaluate recommendation policies in terms of the trade-odd beween relevance and fairness, and propose a recommendation model considering user's tolerance towards fairness. The idea of counterfactual is used not only in fairness model design but also in fairness diagnosis. Specifically, fairness diagnosis aims to find out the reasons that cause model unfairness. Inspired by the idea of counterfactual explanation [130, 210], Ge et al. [137] propose a counterfactual reasoning approach to learn critical features that significantly influence the fairness-utility trade-off and use them as fairness explanation for feature-based recommendation.

9.2.3 Structural Equations

A Structural Causal Model consists of a causal graph, which captures the direct causal relations among variables, and a set of structural equations, which builds the quantitative relations among variables. As we introduced in Section 6, if the structural equations are given, the interventional or counterfactual outcomes can be obtained by replacing the value of variables in structural equations. Inspired by this idea, some works on fairness utilize the learned or pre-defined structural equations. For example, some works [92, 211, 212, 213, 214, 215] model different causal effects from learned structural equations to discover discrimination and further remove them. Kilbertus et al. [204] develop a practical procedure to remove discrimination given the structural equation model.

9.2.4 Causal Graphs

Some other causality-based methods utilize causal graphs to capture the underlying data generation mechanism and apply other techniques to achieve fairness. For example, Huang et al. [216] use the d-separation set identified from the causal graph to design a fair upper confidence bound bandit algorithm for online recommendation. Li et al. [139] design a model based on a causal graph to generate feature independent user representations via adversary learning. Concretely, the model trains a predictor and an adversarial classifier simultaneously, where the predictor learns the representations for recommendation and the classifier minimizes the predictor's ability to predict the sensitive features.

9.3 Open Problems

As we introduced above, researchers start to realize the importance of considering causality-based fairness in recommendation [76, 201]. However, the foundation of causal fairness in recommendation has not been well established. Specifically, the fairness techniques are well explored in classification tasks, however, those techniques may not be directly migrated to the recommendation problem even if the recommendation can be considered as a classification task in some cases. For example, a straightforward method [91] to achieve counterfactual fairness in classification is removing sensitive attributes from the input to guarantee the independence between the outcomes and the sensitive features. However, in recommender systems, some existing approaches do not use features for recommendation, such as most collaborative filtering based models [217], but still suffer from unfairness. The reason is that the interaction information contains the hidden relationship between sensitive features and user-item interaction, and this underlying relationship will be captured by the model during the collaborative learning thus leading to unfairness. Therefore, it is critical to have more explorations about the underlying causal mechanism of unfairness. Additionally, it can help the community to establish a connection between bias and fairness.

10 Causal Robustness in Recommendation

Recently, the robustness of machine learning become increasingly important. Because model time is very time-consuming therefore, the recommender system models are not re-trained frequently in practice. Traditionally, the recommender system assumes that the pattern of the training dataset and the test dataset are the same. However, there is a difference between the training dataset and the real-world data. The difference might be caused by the naturally distribution shift or intend attacking [218].Training on such a training dataset will result in performance decreasing when we apply the model to real-world data. In this case, how to construct a robust model is very important.

10.1 Problem Introduction

To begin with, we need to know which aspects harm the robustness of the recommender system. In general the dataset will be split into three subset in the training progress (training set, validation set and test set). Most of the robustness happen on training set and test set. For example, if the training dataset if not big enough can this may cause the overfitting or underfitting problem. In this case, we may get a bad results on the test set. Specifically, the robustness problem can be categorized as following:

- **Distributional shift** Many exist recommender systems assume that the distribution for the training set and test set are identical. However this assumption do not meet the real-world scenarios, and this makes lots of exist recommendation models cannot achieve the performance we expect when we deploy them online [219]. Many of the current recommender system models are trained based on existing collected datasets. The distribution of the data can be different when the new model is deployed [220]. Even though the data is new collected, there are still transformation risks [218]. Because there is a time period between training the model and deploying the model. It is long enough for the information of the collected be to attacked or changed due to the processing error.
- **Transformation** Most of the recommender system are training with the users' and items' feature. Based on the feature, the recommender system can provide the users a set of recommendation items. However, the users' and items' feature can be corrupted or misleading. For example, if users use VPN when they purchase a item, then the location information could be wrong. With such data, the recommender system may provide wrong items to the users.

- **Attack** Except the transformation, attack may also change the correctness of the training data. With the development of e-commence, more people start to purchase items online. There are huge benefits associated with this. Therefore, the system are highly likely to be attacked with the purpose to increase or decrease the ranking of some items. For example, the users may unwillingly changed their rating and reviews of the purchased product.
- **Sparsity** Compared to the huge number of user and item, most of the sequential recommendation train set is sparsity. As we mentioned above, this may cause overfitting or underfitting problem and lead to low performance on test set.

10.2 Method in Robustness

10.2.1 Counterfactual

Recommendation model suffers from the data sparsity problem. For example, in the e-commerce application, compared to the large number of user and items, the users purchase history is quiet sparsity. Therefore, the model cannot get enough data for training and result in the low prediction performance. One approach to solve this problem is counterfactual data augmentation. By using counterfactual reasoning to generate new training data, and together with the original data can enhance the performance of the recommendation model. Counterfactual Data-Augmentation Sequential Recommendation (CASR) provides us a framework to solve the problem [221]. For a training sample $(\{u, t^1, t^2, ...t^l\}, t^{l+1})$, the model will first indicate an index d, and replace a t^d with an item t^a. Suppose $\mathbf{e}_t \in R^D$ is the embedding of item t, D is the embedding size of the item. For a given sample $(\{u, t^1, t^2, ...t^l\}, t^{l+1})$, the model will optimize the following object:

$$\begin{aligned} &\min_{t^a \in C} \|\mathbf{e}_{t^a} - \mathbf{e}_{t^d}\|_2^2 \\ s.t.\ &t^{l+1} \neq arg \max_{t \in I} \mathcal{S}(t|u, t_1, ..., t^{d-1}, t^a, t^{d+1}, ..., t^l) \end{aligned} \tag{22}$$

where $\mathcal{I}$ is the set of all item.$\mathcal{C}$ is the item set for replacement, which can be specified as $\mathcal{I}$ or other set to involve of some prior knowledge. $\mathcal{S}$ is the sampler used to generate new sequential data. In this function, the object tries to minimize the distance between the original item and the replace item. And the constraint make sure the changed item is not the original one. In this case, we can generate data that not the same as the original one but similar to the original one.

10.2.2 Causal Graph

Some existing works introduced the idea of causal representation learning to mitigate the distributional shift problem. The model split the user features into the observed group and unobserved group and set two types of preference depending on whether it is affected by the observed feature or not [222]. According to the casual graph, a framework is created to model the interaction generation procedure. And to deal with the unobserved feature, they design a new Variational Auto-Encoder (VAE) to infer the unobserved feature from the historical interaction and observed features.

10.2.3 Reweighting

Moreover, reweighting methods have been introduced to improve the robustness of recommendation, for example, Li et al.[113] consider to enhance the robustness of recommendation when there are agnostic distributional shifts between training data and testing data. To this end, the paper introduces a personalized feature selection method for Factorization Machines (FMs) through referring to the confounder balancing approach to balance the confounders of each feature. In specific, considering there is usually no prior knowledge of the causal structure of input variables in FMs, the paper considers to treat every feature as a treatment variable and aims to estimate its causal effect on the outcome. When one feature is treated as a treatment, the other features are considered as confounders. The paper refers to the idea of confounder balancing [223, 224] to learn a weight matrix to reweight each sample through balancing the distributions of confounders across different treatment features, so that FMs will assign a weight to each feature that implies its causal effect on the target variable, and thus help to select causal features for achieving robust recommendations.

10.3 Open Problems

Existing works on robustness problem can only focus on some specific problems. For example, using counterfactual data augmentation problem to mitigate sparsity problem and using causal representation learning to solve distribution shift problem. And if we apply these methods on other robustness problems, the experimental performance might decrease a lot. And most of the current existing works are unexplained. They might have good performance on solving some problems, but they cannot explain which part of the model improves the performance and which part of the model can have more improvement. An explained robustness method that can be applied on multiple problems is a great challenge.

11 Uplift-based Recommendation

Modern recommender systems usually aim to recommend items that users are most likely to interact with (e.g., click, purchase, etc.). However, users may interact with some items even without recommendation. Based on this fact, some existing works propose to recommend items with high interaction probability lift instead of high interaction probability values.

A closely related area is uplift modeling, which refers to the techniques used to estimate the incremental impact of a treatment on the outcomes. Uplift modeling is both a causal inference and a machine learning problem [225]. It is a causal inference problem because the two required outcomes (i.e., receive treatment or no treatment) for calculating the incremental impact are exclusive for an individual. It is also a machine learning problem because it needs models to predict reliable uplift values for decision making. Theoretically, uplift modeling aims to estimate a treatment effect on outcomes [225]. There are three main approaches in existing literature: the Two-Model approach trains two models on treated data and controlled data respectively, and uses the difference between two predictions to calculate the uplift value [226]; the class transformation approach builds the connection between the treated group and the controlled group based on some assumptions [227]; the direct estimation approach designs a model to directly estimate the uplift value [228, 229].

11.1 Problem Introduction

Recommender systems have been employed in several industrial domain to increase the profit of the business and improve user engagement. To achieve this goal, most of the recommendation models are designed to increase user action (e.g., click, purchase, etc.) by recommending items that have highest interaction probabilities. However, most recommender systems neglect a fact that users may take actions on some items regardless of whether the system recommends them [75, 230]. For example, a user will purchase bottled water with 95% probability and energy drink with 50% probability if the system recommends them. In the opinion of most traditional recommender systems, it would be better to recommend bottled water since it is more likely to be purchased by the user. However, if the system does not recommend them, bottled water, as a product for daily use, may still have 90% probability to be purchased, while energy drink may only have 20% probability of being purchased. Recommending energy drink seems to be a better choice since it has higher lift of purchase probability (i.e., 30% vs 5%), which in turn may be expected to lead to more profit. Based on this motivation, there is a trend to design uplift-based recommender systems which aim to recommend items with high lift.

Some previous works have been aware of the impact of recommendation but have not solve it from a causal view. For example, Bodapati [231] proposes a two-stage model which separately trains the awareness and satisfaction stages for items. By training the model based on firm-initiated purchase data (i.e., purchases as a consequence of recommendation) and self-initiated purchase data (i.e., purchases other than firm-initiated purchases), the model aims to recommend items that maximize the expected incremental number of purchase from recommendation. Sato et al. [230] propose a purchase prediction model which incorporates individual differences in recommendation responsiveness. More specifically, the model includes user-specific and item-specific responsiveness to maximize the impact of recommendation.

An uplift in recommender systems is defined as an increase of user actions (e.g., click, purchase, like, etc.) caused by recommendations. Considering the uplift is defined as difference between situations with and without recommendation, from the perspective of causal inference, the uplift can be mathematically represented by potential outcomes. More specifically, taking recommendation as the treatment, let $Y(1)$ be the potential outcome with a recommendation, $Y(0)$ be the potential outcome without a recommendation. Considering the binary situation, $Y(1) = 1$ and $Y(0) = 1$ imply that a user will take actions on the item with and without recommendation, respectively. The uplift of an item for a user is $Y(1) - Y(0)$. In the following subsection, we will introduce some existing works on uplift-based recommendation models with causal inference.

11.2 Causal Methods

11.2.1 Data Processing

One challenge of estimating the uplift value is that each individual cannot observe both the factual and counterfactual outcomes (i.e., outcomes with and without recommendations). Thus there is no observed ground truth for the uplift value (i.e., the causal effect of recommendation). To overcome this issue, one possible solution is regarding the training data. Sato et al. [75] propose a sampling method on the observational data for an uplift-based optimization. Specifically, by observing purchase and recommendation logs, for a given user, an item can be either purchased or not purchased and either recommended or not recommended. The proposed optimization samples positive and negative instances that are specific to the uplift task from four classes items (i.e., recommended and purchased, recommended and not purchased, not recommended and purchased, not recommended and not purchased). Therefore, by taking the sampled labels for uplift task as the ground truth, the proposed optimization is able to learn the uplift value for user-item pairs. Except for the sampling methods on observational data, training on experimental data is also an available option. Shang et al. [86] propose a reinforcement learning based approach, which incorporate a deep uplift network to learn the causal effect of different actions as a reward function. The uplift network learns from the training data collected from a randomized experiment.

11.2.2 Counterfactual

According to the calculation of the uplift value, a straightforward way is to estimate the counterfactual outcomes. Although the randomized experiment is ideal for estimating the causal effect, it is impractical to apply randomized experiment in all recommendation scenarios since it is time-consuming and expensive. Therefore, it is essential to estimate the counterfactual outcomes only based on the observational data. Inspired by the idea of collaborative filtering, Xie et al. [87] believe that similar users have both similar tastes on items and similar treatment effect under recommendations. The proposed approach is designed based on tensor factorization with three dimensions as user, item and treatment. More specifically, for a three dimentional tensor with m users, n items and l treatments, the element $y_{u,i,t}$ can be predicted as follows.

$$\hat{y}_{u,i,t} = p_u^T q_i + p_u^T d_t + q_i^T d_t \tag{23}$$

where p_u, q_i, d_t are latent representation of user u, item i and treatment t, respectively. The predicted value of $\hat{y}_{u,i,t}$ is used to infer the potential outcome for a user-item pair (u, i) under treatment t. Taking binary treatment setting as an example, the uplift value for a user-item pair (u, i) can be estimated by $\hat{y}_{u,i,t=1} - \hat{y}_{u,i,t=0}$. Sato et al. [88] apply a matching estimator [232] to estimate unobserved counterfactual outcomes and further estimate the causal effect for recommendation. More specifically, following the neighborhood methods in recommender systems, the proposed approach replaces the potential outcomes with the weighted average over the observed outcome for a set of neighbors to calculate the causal effect, where the neighbors can be neighborhood users or neighborhood items.

11.2.3 Reweighting

Estimating causal effect from the observational data only is challenging, since the ground truth is unobservable and the estimation is prone to the biases in the observational data. To overcome this issue, some existing works design IPS-based approaches to estimate unbiased causal effect for recommendation or evaluation. The unbiased estimation of the uplift value (i.e., the causal effect) can be formuted by IPS [233]. In practice, IPS is prone to suffer from high variance issue. To tackle this problem, Sato et al. [233] apply capped inverse propensity scoring (CIPS) to train an unbiased uplift-based model; Sato et al. [75] propose a unbiased estimator for uplift-based evaluation using self-normalized inverse propensity scoring (SNIPS) [234]; Xiao and Wang [235] apply doubly robust technique [107, 236] to train an unbiased and robust model for uplift-based recommendation.

11.3 Open Problems

Existing works on uplift-based recommendation mainly focus on representing uplift value and estimating the causal effect using potential outcome framework. Structural causal model, as a power tool for causal inference, has rarely been used for uplift-based recommendation. Existing works using structural causal models are trying to estimate user's preference if the system recommends a certain item, which can be estimated by *do*-operations on designed causal graph. However, it is still not clear how to estimate the preference without recommendation using *do*-operation. First, the structural causal model requires the designed causal graph. Existing works on causal recommendation using structural causal models rarely explicitly involve the impact of recommendation into the causal graph, however for uplift-based recommendation, whether requiring a specific causal graph that explicitly depict the impact of recommendation still needs to be discussed. Secondly, mathematical representation of the preference without recommendation using *do*-operation is also a challenge. Finally, for uplift-based recommendation, if the designed causal graph and mathematical representation of preference without recommendation are decided, applying causal techniques on preference without recommendation may differ from existing works.

12 Causal Unbiasedness in Recommendation

Nowadays, recommendation algorithms have been widely used in several applications to alleviate information overloading in our daily life. Although recommender systems (RS) have obtained huge impacts in a wide range of real-world applications, it still faces many bias issues which are challenging, and if left unattended, will affect the long-term benefits of the recommender systems. Bias issues are common in RS since one nature of RS is the feedback loop. Following a generally accepted understanding [35, 36], the feedback loop in RS can be divided into three parts from a bird's-eye view: 1) the data collection part (user → data); 2) the model training part (data → model); 3) the model serving part (model → user). Different definitions of bias issues exist in each part and the whole feedback loop. We will introduce more details in the following parts.

12.1 Problem Introduction

As we mentioned, the bias issues exist in each part as well as the whole of the feedback loop. We will introduce the different definition of bias in the feedback loop of RS as follows:

- **Bias in Data** refers to the distribution difference between the collected data for training and the ideal test data. Typically, the training data for RS is observational instead of experimental. The user decision may be affected by several factors such as exposure mechanism of RS, thus the training distribution is different from the test distribution. Additionally, the training data may not truly represent user preference, misleading recommendation model to inaccurate prediction. We will introduce four kinds of bias in data as follows:
  - *Selection Bias*: Selection bias stems from users' explicit feedback (i.e., ratings). Selection bias means the observed ratings are not representative of all ratings due to users' selection. It is also referred as missing-not-at-random (MNAR).
  - *Exposure Bias*: Exposure bias usually happens in recommendations with implicit feedback. Since the information about which item the user dislikes is unavailable in observed data, the learning process will use unobserved interactions to represent negative preference. Exposure bias means unobserved interactions do not necessarily represent the user's negative preference since the users are merely exposed to a small portion of items.
  - *Conformity Bias*: Conformity Bias means that users tend to behave similarly to the others in the group, even if their behavior goes against their own judgment, which makes the feedback may not represent users' true preference.
  - *Position Bias*: Position bias is common in recommendation, especially the results are presented by a ranking list. Position bias means that users tend to interact with items in higher position in the recommendation list even if the items in higher position may not be highly relevant.
- **Bias in Model** refers to the bias in the model design. Bias i not always harmful. In fact, the bias in model empower the model to achieve the ability to generalize the prediction to unobserved examples.
  - *Inductive Bias*: Inductive bias represents the assumptions made the model designer to better learn the objective and to generalize beyond training data.
- **Bias in Results** refers to the phenomenon that the recommendation algorithms tend to exhibit bias in recommendation results presented to users. Typically, the biases in recommendation results are studied from two perspectives, one is popularity bias and the other is unfairness. We have introduced fairness and related methods in section 9, thus in this section, we will limit the bias in results to popularity bias.
  - *Popularity Bias*: Popularity bias refers to the phenomenon that popular items are recommended more frequently than their popularity warrant.

- **Feedback Loop Bias** refers to the amplified bias introduced by the whole RS feedback loop mechanism. Data bias will lead to data imbalance and result in bias issues in recommendation results, while the biased recommendation will in turn impact the user's behavior and further amplify the bias in the future recommendation. Taking popularity bias as an example, the popular items get more exposure in the observed data, which in turn obtain increase opportunity to be recommended, resulting in amplified bias, where popular items become more popular and non-popular items become even less popular [237, 238, 239]. These amplified bias caused by feedback loop, if left unattended, will result in echo chambers [115, 240] or filter bubble [241, 242, 243, 244], which will decrease the diversity and increase the homogenization.

In general, there are two ways for debiasing in recommender systems, one is debiasing during training and the other is debising during evaluation. Introducing causal inference into debias recommendation makes a great success in recent years. In the following part, we will introduce existing works on debiased recommendation models based on causal inference.

12.2 Causal Methods

12.2.1 Data Processing

To address the bias problem in recommender systems, one straightforward solution is to leverage unbiased data [82, 96, 97, 98, 99, 100, 101, 102]. As we mentioned in section, combining the limited experimental data and observational data is a possible solution under the relaxed ignorability assumption. In recommender systems, the experimental data, which is also called as unbiased data, intervene the system by using a random recommendation policy instead of a normal recommendation policy. More specifically, for each user, they do not use recommendation models to show items, but instead randomly select some items to show. Leveraging unbiased data helps to achieve debiased prediction because applying random recommendation will break the feedback loop. The key challenge is how to incorporate a small portion of unbiased data into model design. For example, Rosenfeld et al. [96] and Bonner and Vasile [82] apply two recommendation models for biased data and unbiased data respectively and connect two models by regularization. Yuan et al. [97] learn a imputation models with unbiased data for ad click prediction. Chen et al. [101] leverage unbiased data by meta-learning. Despite the effectiveness on handle biases by using unbiased data, collecting unbiased data will randomly recommend items to users instead of using personalized recommendation model, which will inevitably hurt users' experience and revenues of the platform.

12.2.2 Reweighting

Another commonly used method is based on reweighting, which use inverse propensity scores to reweight the data sample for different bias issues, such as selection bias [90, 106], exposure bias [70, 72, 73, 245], position bias [246, 247, 248, 249, 250, 251], etc. The key challenge is how to estimate the propensity scores and how to apply it into optimization. Some works [70, 252] use popularity-based propensity estimator. Some works [73, 248, 250, 251] propose a dual problem to both optimize a propensity estimator and a recommendation model. Some works [249] propose to learn propensity scores from the observational data. Some works [207, 235] use doubly robust model to handle inaccurate propensity estimators. Inverse propensity scoring methods have some limitations, such as inaccurate propensity scores and suffering from high variance problem [206].

12.2.3 Causal Adjustment

Causal adjustment is another promising direction for addressing bias issues [77, 79, 81, 114, 115]. With the help of do-operator, the designed models aim to estimate the causal preference $P(Y|U, do(V))$ with intervening item exposure rather than the pure associative preference $P(Y|U, V)$ estimated by traditional recommendation models. Intuitively, it can be understood as to answer a counterfactual question: what would the preference be if we intervene to expose the item to the user? Causal adjustment is used to estimate the causal preference with observational data. More specifically, causal adjustment includes back-door adjustment [68], front-door adjustment [68], etc. Based on the designed causal graph representing the underlying mechanism of data generation in recommender systems, the first thing is to identify a set of variables satisfying the corresponding criterion (e.g., back-door criterion for back-door adjustment, front-door criterion for front-door adjustment), then apply causal adjustment on identified variable set to estimate the causal preference. For example, Zhang et al. [79] apply back-door adjustment to mitigate the exposure bias caused by the item popularity; Xu et al. [77] leverage front-door adjustment to remove the effect of unobserved confounders; Wang et al. [114] utilize back-door adjustment to mitigate the effect of popularity bias. Causal adjustment requires to identify a set of variables satisfying the corresponding criterion, however, given a reasonable causal graph for recommender systems, it is not always find out a set of variables satisfying such criterion. But the designed causal graph will guide the model design from other ways.

12.2.4 Causal Graph

Causal graph, as an effective and powerful tool for causal modeling, is used to depict the data generation process in recommender systems. Based on the designed causal graph, researchers will take it as the guidance to design causal models for debiasing [80, 253]. For example, Zhao et al. [253] and Zheng et al. [80] disentangle the effect from bias and user's preference based on the designed causal graph and recommend items solely based on user's preference; Wei et al. [95] and Wang et al. [94] represent the counterfactual world based on the designed causal graph and perform counterfactual reasoning for recommendation.

12.3 Open Problems

Inverse propensity scoring (IPS) is a valuable method for debiasing. However, the effectiveness of IPS methods highly rely on the correctness of propensity scores. How to obtain correct propensity scores is still an important yet unsolved question. Existing work usually design simple propensity estimator based on some item characteristics, such as popularity-based propensity [70], or learn the propensity scores from data [73, 248, 249, 250, 251]. Whether using correct propensity scores can be only estimated indirectly through the improvement for recommendation performance. Therefore, quantitative evaluation of the correctness of propensity scores is still an open problem and need further exploration.

13 Open Problems and Future Directions

13.1 Underlying Causal Mechanisms

Recall the existing works we introduced above, most of them are based on the underlying causal mechanisms of recommender systems, which are represented by pre-defined causal relations. In general, there are three levels of pre-defined causal relations. The first level is identifying cause and effect only. For example, IPS methods only investigate the quantitative relationship between two variable, one is cause, the other is effect. For example, in some IPS based models for debiasing in recommendation, the cause is item exposure, and the effect is the probability of interactions. The second one is defining causal graphs, which identify the causal relationship between all variable pairs (i.e., whether causal relation exists, the direction of causal relation if exists). By pre-defining the causal graph, some existing works design models with the guidance of causal graph. For example, some works [80, 253] disentangle multiple cause on the effect based on the defined causal graph to achieve unbiasedness. The last level is structural causal models, which define not only the causal relations but also quantitative relations (i.e., structural equations take causes as input and return the value of effect). The effectiveness of proposed models are highly related to the correctness of underlying causal mechanism. Currently, most of the existing works define the underlying causal mechanism through expert knowledge. The correctness of the pre-defined causal mechanism can only be indirectly reflected by the recommendation performance. Therefore, a direct and quantitative evaluation of the defined causal mechanisms deserves for further exploration. Another observation is that different models may have different pre-defined causal mechanisms even under the same practical scenario. As such, we believe that a universal causal mechanism should be proposed.

13.2 Causal Discovery

Apart from concerns about the accuracy of pre-defined causal mechanisms, another limitation of pre-defined causal mechanism is that pre-defined causal mechanisms by expert knowledge are usually quite simple, which only involve few factors into consideration. However, in real-world scenario, the decision-making process (i.e., the underlying causal mechanism of recommender systems) may involve much more factors, which beyond the comprehension of domain experts. Therefore, learning causal relations from data is an important yet unsolved problem in recommender systems. There exist few works [125, 126] design causal discovery methods based on continuous optimization [127, 254] for recommendation. The learned causal mechanism can be used for explainable recommendation or be leveraged to improve recommendation. Therefore, it is a promising direction to propose causal discovery methods for recommendation. It is also a challenge to evaluate the proposed causal discovery methods for recommendation. Since there is no ground-truth causal mechanism in real-world data, the causal discovery methods in recommendation are usually indirectly evaluated by recommendation performance. To directly evaluate causal discovery methods for recommendation, one possible solution is using simulation (we will introduce it in the next section).

13.3 Causality Driven Simulations

Simulation is one of the most powerful approach to build environments in which the recommender systems can be measured and analyzed. Building simulation for recommendation will benefit both industry and academia. For example, for industry, simulation provide controllable environment for practitioners to analyze the objectives of interest, such as some business purpose, to accelerate the pace of application development without the ethical risks. For researchers in academia, due to the restrictions of accessibility of real-world recommender systems, some proposed methods cannot be evaluated. This issue can be addressed by using simulations. Existing simulations leverage reinforcement learning techniques to simulate the decision making process under a designed environment. However, existing simulations without underlying causal mechanisms may lead to inaccurate and unstable decision-making. Leveraging causal mechanisms into simulation will achieve more stable system for long-term analysis and causal-related analysis as well. For example, causality driven simulations can be used to evaluate causal discovery methods in recommendation. Thus, causality driven simulation will play an essential role in recommender systems, which deserves further explorations.

14 Conclusion

In this survey, we provide comprehensive review of causal inference methods for recommendation. We first provide the fundamental knowledge of recommender systems. We then introduce existing work in perspective of both causal inference and recommender systems. More specifically, on the one hand, we introduce knowledge about causal inference and demonstrate its connection with recommender systems, on the other hand, we introduce different problems in recommender systems and how causal inference applied. Finally, we further list some open problems and future directions. We hope this survey can benefit researchers and practitioners in this area and inspire more research work in causal inference for recommendation.

References

[1] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender systems: an introduction*. Cambridge University Press, 2010.

[2] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM Computing Surveys (CSUR)*, vol. 52, no. 1, pp. 1–38, 2019.

[3] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE transactions on knowledge and data engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[4] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "Grouplens: an open architecture for collaborative filtering of netnews," in *CSCW*, 1994, pp. 175–186.

[5] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, "Grouplens: applying collaborative filtering to usenet news," *Communications of the ACM*, vol. 40, no. 3, pp. 77–87, 1997.

[6] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *WWW*, 2001, pp. 285–295.

[7] G. Linden, B. Smith, and J. York, "Amazon. com recommendations: Item-to-item collaborative filtering," *IEEE Internet computing*, vol. 7, no. 1, 2003.

[8] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[9] A. Mnih and R. R. Salakhutdinov, "Probabilistic matrix factorization," in *Advances in neural information processing systems*, 2008, pp. 1257–1264.

[10] S. Rendle, "Factorization machines," in *2010 IEEE International Conference on Data Mining*. IEEE, 2010, pp. 995–1000.

[11] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir *et al.*, "Wide & deep learning for recommender systems," in *Proceedings of the 1st workshop on deep learning for recommender systems*, 2016, pp. 7–10.

[12] H.-J. Xue, X. Dai, J. Zhang, S. Huang, and J. Chen, "Deep matrix factorization models for recommender systems." in *IJCAI*, vol. 17. Melbourne, Australia, 2017, pp. 3203–3209.

[13] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *WWW*, 2017, pp. 173–182.

[14] C.-K. Hsieh, L. Yang, Y. Cui, T.-Y. Lin, S. Belongie, and D. Estrin, "Collaborative metric learning," in *In WWW*, 2017, pp. 193–201.

[15] L. Zheng, V. Noroozi, and P. S. Yu, "Joint deep modeling of users and items using reviews for recommendation," in *WSDM*, 2017.

[16] Y. Zhang, Q. Ai, X. Chen, and W. B. Croft, "Joint representation learning for top-n recommendation with heterogeneous information sources," in *CIKM*, 2017, pp. 1449–1458.

[17] F. Zhang, N. J. Yuan, D. Lian, X. Xie, and W.-Y. Ma, "Collaborative knowledge base embedding for recommender systems," in *KDD*, 2016, pp. 353–362.

[18] Q. Ai, V. Azizi, X. Chen, and Y. Zhang, "Learning heterogeneous knowledge base embeddings for explainable recommendation," *Algorithms*, vol. 11, no. 9, p. 137, 2018.

[19] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-based recommendations on styles and substitutes," in *SIGIR*. ACM, 2015.

[20] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," in *ICLR*, 2016.

[21] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018.

[22] M. Kompan and M. Bieliková, "Content-based news recommendation," in *International conference on electronic commerce and web technologies*. Springer, 2010, pp. 61–72.

[23] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in *Proceedings of the 10th ACM conference on recommender systems*, 2016, pp. 191–198.

[24] Q. Liu, S. Wu, and L. Wang, "Deepstyle: Learning user preferences for visual recommendation," in *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, 2017, pp. 841–844.

[25] R. Burke, "Hybrid recommender systems: Survey and experiments," *User modeling and user-adapted interaction*, vol. 12, no. 4, pp. 331–370, 2002.

[26] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[27] J. Pearl, *Causality*. Cambridge university press, 2009.

[28] ——, "Causal inference in statistics: An overview," *Statistics surveys*, vol. 3, pp. 96–146, 2009.

[29] Y. Zhang, X. Chen *et al.*, "Explainable recommendation: A survey and new perspectives," *Foundations and Trends® in Information Retrieval*, vol. 14, no. 1, pp. 1–101, 2020.

[30] X. Chen, Y. Zhang, and J.-R. Wen, "Measuring" why" in recommender systems: a comprehensive survey on the evaluation of explainable recommendation," *arXiv preprint arXiv:2202.06466*, 2022.

[31] Y. Li, H. Chen, S. Xu, Y. Ge, J. Tan, S. Liu, and Y. Zhang, "Fairness in recommendation: A survey," *arXiv preprint arXiv:2205.13619*, 2022.

[32] Y. Wang, W. Ma, M. Zhang*, Y. Liu, and S. Ma, "A survey on the fairness of recommender systems," *ACM Journal of the ACM (JACM)*, 2022.

[33] Y. Ge, S. Liu, Z. Fu, J. Tan, Z. Li, S. Xu, Y. Li, Y. Xian, and Y. Zhang, "A survey on trustworthy recommender systems," *arXiv preprint arXiv:2207.12515*, 2022.

[34] S. Wang, X. Zhang, Y. Wang, H. Liu, and F. Ricci, "Trustworthy recommender systems," *arXiv preprint arXiv:2208.06265*, 2022.

[35] W. Fan, X. Zhao, X. Chen, J. Su, J. Gao, L. Wang, Q. Liu, Y. Wang, H. Xu, L. Chen *et al.*, "A comprehensive survey on trustworthy recommender systems," *arXiv preprint arXiv:2209.10117*, 2022.

[36] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, and X. He, "Bias and debias in recommender system: A survey and future directions," *arXiv preprint arXiv:2010.03240*, 2020.

[37] H. Ko, S. Lee, Y. Park, and A. Choi, "A survey of recommendation systems: recommendation models, techniques, and application fields," *Electronics*, vol. 11, no. 1, p. 141, 2022.

[38] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang,

"A survey on causal inference," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 5, pp. 1–46, 2021.

[39] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *ACM Computing Surveys (CSUR)*, vol. 53, no. 4, pp. 1–37, 2020.

[40] M. J. Vowels, N. C. Camgoz, and R. Bowden, "D'ya like dags? a survey on structure learning and causal discovery," *ACM Computing Surveys (CSUR)*, 2021.

[41] C. Gao, Y. Zheng, W. Wang, F. Feng, X. He, and Y. Li, "Causal inference in recommender systems: A survey and future directions," *arXiv preprint arXiv:2208.12397*, 2022.

[42] A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver, "Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering," in *Proceedings of the fourth ACM conference on Recommender systems*, 2010, pp. 79–86.

[43] R. Van Meteren and M. Van Someren, "Using content-based filtering for recommendation," in *Proceedings of the machine learning in the new information age: MLnet/ECML2000 workshop*, vol. 30, 2000, pp. 47–56.

[44] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in *The adaptive web*. Springer, 2007, pp. 325–341.

[45] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," *Knowledge-based systems*, vol. 46, pp. 109–132, 2013.

[46] T. Alashkar, S. Jiang, S. Wang, and Y. Fu, "Examples-rules guided deep neural network for makeup recommendation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[47] D. Liang, M. Zhan, and D. P. Ellis, "Content-aware collaborative music recommendation using pre-trained neural networks." in *ISMIR*, 2015, pp. 295–301.

[48] R. Burke, "Hybrid web recommender systems," in *The adaptive web*. Springer, 2007, pp. 377–408.

[49] E. Çano and M. Morisio, "Hybrid recommender systems: A systematic literature review," *Intelligent Data Analysis*, vol. 21, no. 6, pp. 1487–1524, 2017.

[50] M. Balabanović and Y. Shoham, "Fab: content-based, collaborative recommendation," *Communications of the ACM*, vol. 40, no. 3, pp. 66–72, 1997.

[51] P. Melville, R. J. Mooney, R. Nagarajan *et al.*, "Content-boosted collaborative filtering for improved recommendations," *Aaai/iaai*, vol. 23, pp. 187–192, 2002.

[52] H. Fang, D. Zhang, Y. Shu, and G. Guo, "Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations," *ACM Transactions on Information Systems (TOIS)*, vol. 39, no. 1, pp. 1–42, 2020.

[53] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme, "Factorizing personalized markov chains for next-basket recommendation," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 811–820.

[54] H. Hu, X. He, J. Gao, and Z.-L. Zhang, "Modeling personalized item frequency information for next-basket recommendation," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1071–1080.

[55] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," *arXiv preprint arXiv:1511.06939*, 2015.

[56] C.-Y. Wu, A. Ahmed, A. Beutel, A. J. Smola, and H. Jing, "Recurrent recommender networks," in *Proceedings of the tenth ACM international conference on web search and data mining*, 2017, pp. 495–503.

[57] J. Tang and K. Wang, "Personalized top-n sequential recommendation via convolutional sequence embedding," in *Proceedings of the eleventh ACM international conference on web search and data mining*, 2018, pp. 565–573.

[58] F. Yuan, A. Karatzoglou, I. Arapakis, J. M. Jose, and X. He, "A simple convolutional generative network for next item recommendation," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 582–590.

[59] S. Wang, L. Hu, L. Cao, X. Huang, D. Lian, and W. Liu, "Attention-based transactional context embedding for next-item recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[60] X. Chen, H. Xu, Y. Zhang, J. Tang, Y. Cao, Z. Qin, and H. Zha, "Sequential recommendation with user memory networks," in *Proceedings of the eleventh ACM international conference on web search and data mining*, 2018, pp. 108–116.

[61] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer." *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[62] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[63] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[64] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311*, 2022.

[65] S. Geng, S. Liu, Z. Fu, Y. Ge, and Y. Zhang, "Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5)," in *Proceedings of the 16th ACM Conference on Recommender Systems*, 2022.

[66] Z. Zhu, X. Hu, and J. Caverlee, "Fairness-aware tensor-based recommendation," in *Proceedings of the 27th ACM international conference on information and knowledge management*, 2018, pp. 1153–1162.

[67] P. W. Holland, "Statistics and causal inference," *Journal of the American statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.

[68] M. Glymour, J. Pearl, and N. P. Jewell, *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[69] B. B. Frey, *The SAGE encyclopedia of educational research, measurement, and evaluation*. Sage Publications, 2018.

[70] Y. Saito, S. Yaginuma, Y. Nishino, H. Sakata, and K. Nakata, "Unbiased recommender learning from missing-not-at-random implicit feedback," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, pp. 501–509.

[71] Y. Saito, "Unbiased pairwise learning from implicit feedback," in *NeurIPS 2019 Workshop on Causal Machine Learning*, 2019.

[72] L. Yang, Y. Cui, Y. Xuan, C. Wang, S. Belongie, and D. Estrin, "Unbiased offline recommender evaluation for missing-not-at-random implicit feedback," in *Proceedings of the 12th ACM conference on recommender systems*, 2018, pp. 279–287.

[73] Z. Zhu, Y. He, Y. Zhang, and J. Caverlee, "Unbiased implicit recommendation and propensity estimation via combinational joint learning," in *Fourteenth ACM Conference on Recommender Systems*, 2020, pp. 551–556.

[74] S. Ding, P. Wu, F. Feng, Y. Wang, X. He, Y. Liao, and Y. Zhang, "Addressing unmeasured confounder for recommendation with sensitivity analysis," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 305–315.

[75] M. Sato, J. Singh, S. Takemori, T. Sonoda, Q. Zhang, and T. Ohkuma, "Uplift-based evaluation and optimization of recommenders," in *Proceedings of the 13th ACM Conference on Recommender Systems*, 2019, pp. 296–304.

[76] K. Makhlouf, S. Zhioua, and C. Palamidessi, "Survey on causal-based machine learning fairness notions," *arXiv preprint arXiv:2010.09553*, 2020.

[77] S. Xu, J. Tan, S. Heinecke, J. Li, and Y. Zhang, "Deconfounded causal collaborative filtering," *arXiv preprint arXiv:2110.07122*, 2021.

[78] Y. Wang, D. Liang, L. Charlin, and D. M. Blei, "Causal inference for recommender systems," in *Fourteenth ACM Conference on Recommender Systems*, 2020, pp. 426–431.

[79] Y. Zhang, F. Feng, X. He, T. Wei, C. Song, G. Ling, and Y. Zhang, "Causal intervention for leveraging popularity bias in recommendation," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 11–20.

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

[81] S. Xu, Y. Ge, Y. Li, Z. Fu, X. Chen, and Y. Zhang, "Causal collaborative filtering," *arXiv preprint arXiv:2102.01868*, 2021.

[82] S. Bonner and F. Vasile, "Causal embeddings for recommendation," in *Proceedings of the 12th ACM conference on recommender systems*, 2018, pp. 104–112.

[83] J. Pearl, "The mathematics of causal inference," *Forthcoming, IMS*, 2013.

[84] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3076–3085.

[85] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *International conference on machine learning*. PMLR, 2016, pp. 3020–3029.

[86] W. Shang, Q. Li, Z. Qin, Y. Yu, Y. Meng, and J. Ye, "Partially observable environment estimation with uplift inference for reinforcement learning based recommendation," *Machine Learning*, vol. 110, no. 9, pp. 2603–2640, 2021.

[87] X. Xie, Z. Liu, S. Wu, F. Sun, C. Liu, J. Chen, J. Gao, B. Cui, and B. Ding, "Causcf: Causal collaborative filtering for recommendation effect estimation," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 4253–4263.

[88] M. Sato, J. Singh, S. Takemori, and Q. Zhang, "Causality-aware neighborhood methods for recommender systems," in *European Conference on Information Retrieval*. Springer, 2021, pp. 603–618.

[89] S. Xu, Y. Li, S. Liu, Z. Fu, Y. Ge, X. Chen, and Y. Zhang, "Learning causal explanations for recommendation," in *The 1st International Workshop on Causality in Search and Recommendation*, 2021.

[90] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims, "Recommendations as treatments: Debiasing learning and evaluation," in *international conference on machine learning*. PMLR, 2016, pp. 1670–1679.

[91] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *Advances in neural information processing systems*, vol. 30, 2017.

[92] Y. Wu, L. Zhang, and X. Wu, "On discrimination discovery and removal in ranked data using causal graph," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2536–2544.

[93] J. Zhang and E. Bareinboim, "Fairness in decision-making—the causal explanation formula," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[94] W. Wang, F. Feng, X. He, H. Zhang, and T.-S. Chua, "Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1288–1297.

[95] T. Wei, F. Feng, J. Chen, Z. Wu, J. Yi, and X. He, "Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1791–1800.

[96] N. Rosenfeld, Y. Mansour, and E. Yom-Tov, "Predicting counterfactuals from large historical data and small randomized trials," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 602–609.

[97] B. Yuan, J.-Y. Hsia, M.-Y. Yang, H. Zhu, C.-Y. Chang, Z. Dong, and C.-J. Lin, "Improving ad click prediction

by considering non-displayed events," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 329–338.

[98] D. Liu, P. Cheng, Z. Dong, X. He, W. Pan, and Z. Ming, "A general knowledge distillation framework for counterfactual recommendation via uniform data," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 831–840.

[99] R. Jiang, S. Chiappa, T. Lattimore, A. György, and P. Kohli, "Degenerate feedback loops in recommender systems," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 383–390.

[100] J. Yu, H. Zhu, C.-Y. Chang, X. Feng, B. Yuan, X. He, and Z. Dong, "Influence function for unbiased recommendation," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1929–1932.

[101] J. Chen, H. Dong, Y. Qiu, X. He, X. Xin, L. Chen, G. Lin, and K. Yang, "Autodebias: Learning to debias for recommendation," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 21–30.

[102] Z. Lin, D. Liu, W. Pan, and Z. Ming, "Transfer learning in collaborative recommendation for bias reduction," in *Fifteenth ACM Conference on Recommender Systems*, 2021, pp. 736–740.

[103] C. Macdonald, "The simpson's paradox in the offline evaluation of recommendation systems," *ACM Transactions on Information Systems*, 2021.

[104] J. K. Lunceford and M. Davidian, "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study," *Statistics in medicine*, vol. 23, no. 19, pp. 2937–2960, 2004.

[105] K. Hirano, G. W. Imbens, and G. Ridder, "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, vol. 71, no. 4, pp. 1161–1189, 2003.

[106] X. Wang, R. Zhang, Y. Sun, and J. Qi, "Combating selection biases in recommender systems with a few unbiased ratings," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021, pp. 427–435.

[107] H. Bang and J. M. Robins, "Doubly robust estimation in missing data and causal inference models," *Biometrics*, vol. 61, no. 4, pp. 962–973, 2005.

[108] B. K. Lee, J. Lessler, and E. A. Stuart, "Weight trimming and propensity score weighting," *PloS one*, vol. 6, no. 3, p. e18174, 2011.

[109] J. Hainmueller, "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies," *Political analysis*, vol. 20, no. 1, pp. 25–46, 2012.

[110] S. Athey, G. W. Imbens, and S. Wager, "Approximate residual balancing: debiased inference of average treatment effects in high dimensions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 80, no. 4, pp. 597–623, 2018.

[111] K. Kuang, P. Cui, B. Li, M. Jiang, Y. Wang, F. Wu, and S. Yang, "Treatment effect estimation via differentiated confounder balancing and regression," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 14, no. 1, pp. 1–25, 2019.

[112] Z. Shen, P. Cui, T. Zhang, and K. Kunag, "Stable learning via sample reweighting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5692–5699.

[113] Y. Li, H. Chen, J. Tan, and Y. Zhang, "Causal factorization machine for robust recommendation," in *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, 2022, pp. 1–9.

[114] W. Wang, F. Feng, X. He, X. Wang, and T.-S. Chua, "Deconfounded recommendation for alleviating bias amplification," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1717–1725.

[115] S. Xu, J. Tan, Z. Fu, J. Ji, S. Heinecke, and Y. Zhang, "Dynamic causal collaborative filtering," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 2301–2310.

[116] R. Guo, J. Li, and H. Liu, "Learning individual causal effects from networked observational data," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, pp. 232–240.

[117] N. Kallus, A. M. Puli, and U. Shalit, "Removing hidden confounding by experimental grounding," *Advances in neural information processing systems*, vol. 31, 2018.

[118] X. Zhu, Y. Zhang, F. Feng, X. Yang, D. Wang, and X. He, "Mitigating hidden confounding effects for causal recommendation," *arXiv preprint arXiv:2205.07499*, 2022.

[119] J. D. Angrist and J.-S. Pischke, *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

[120] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy, "Deep iv: A flexible approach for counterfactual prediction," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1414–1423.

[121] Z. Si, X. Han, X. Zhang, J. Xu, Y. Yin, Y. Song, and J.-R. Wen, "A model-agnostic causal learning framework for recommendation using search data," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 224–233.

[122] Y. Bengio, T. Deleu, N. Rahaman, R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pal, "A meta-transfer objective for learning to disentangle causal mechanisms," *arXiv preprint arXiv:1901.10912*, 2019.

[123] D. Malinsky and D. Danks, "Causal discovery algorithms: A practical guide," *Philosophy Compass*, vol. 13, no. 1, p. e12470, 2018.

[124] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, "A linear non-gaussian acyclic model for causal discovery." *Journal of Machine Learning Research*, vol. 7, no. 10, 2006.

[125] Z. Wang, X. Chen, Z. Dong, Q. Dai, and J.-R. Wen, "Sequential recommendation with causal behavior discovery," *arXiv preprint arXiv:2204.00216*, 2022.

[126] S. Xu, D. Xu, E. Korpeoglu, S. Kumar, S. Guo, K. Achan, and Y. Zhang, "Causal structure learning with recommendation system," *arXiv preprint arXiv:2210.10256*, 2022.

[127] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P.

Xing, "Dags with no tears: Continuous optimization for structure learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[128] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma, "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014, pp. 83–92.

[129] G. Peake and J. Wang, "Explanation mining: Post hoc interpretability of latent factor models for recommendation systems," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2060–2069.

[130] J. Tan, S. Xu, Y. Ge, Y. Li, X. Chen, and Y. Zhang, "Counterfactual explainable recommendation," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 1784–1793.

[131] X. Wang, Y. Chen, J. Yang, L. Wu, Z. Wu, and X. Xie, "A reinforcement learning framework for explainable recommendation," in *2018 IEEE international conference on data mining (ICDM)*. IEEE, 2018, pp. 587–596.

[132] L. Li, Y. Zhang, and L. Chen, "Personalized transformer for explainable recommendation," 2021.

[133] S. Geng, Z. Fu, Y. Ge, L. Li, G. de Melo, and Y. Zhang, "Improving personalized explanation generation through visualization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 244–255.

[134] Y. Xian, Z. Fu, S. Muthukrishnan, G. De Melo, and Y. Zhang, "Reinforcement knowledge graph reasoning for explainable recommendation," in *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, 2019, pp. 285–294.

[135] Z. Chen, X. Wang, X. Xie, M. Parsana, A. Soni, X. Ao, and E. Chen, "Towards explainable conversational recommendation," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 2994–3000.

[136] A. Ghazimatin, O. Balalau, R. Saha Roy, and G. Weikum, "Prince: Provider-side interpretability with counterfactual explanations in recommender systems," in *WSDM*, 2020, pp. 196–204.

[137] Y. Ge, J. Tan, Y. Zhu, Y. Xia, J. Luo, S. Liu, Z. Fu, S. Geng, Z. Li, and Y. Zhang, "Explainable fairness in recommendation," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.

[138] K. H. Tran, A. Ghazimatin, and R. Saha Roy, "Counterfactual explanations for neural recommenders," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1627–1631.

[139] Y. Li, H. Chen, S. Xu, Y. Ge, and Y. Zhang, "Towards personalized fairness based on causal notion," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1054–1063.

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

[141] Y. Li, Y. Ge, and Y. Zhang, "Tutorial on fairness of machine learning in recommender systems," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 2654–2657.

[142] M. Mansoury, H. Abdollahpouri, J. Smith, A. Dehpanah, M. Pechenizkiy, and B. Mobasher, "Investigating potential factors associated with gender discrimination in collaborative recommender systems," in *The Thirty-Third International Flairs Conference*, 2020.

[143] A. D'Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, and 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.

[144] H. Abdollahpouri and R. Burke, "Multi-stakeholder recommendation and its connection to multi-sided fairness," in *Proceedings of the RMSE workshop held in conjunction with the 13th ACM Conference on Recommender Systems*, 2019.

[145] Y. Ge, S. Liu, R. Gao, Y. Xian, Y. Li, X. Zhao, C. Pei, F. Sun, J. Ge, W. Ou *et al.*, "Towards long-term fairness in recommendation," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021, pp. 445–453.

[146] M.-I. Dascalu, C.-N. Bodea, M. N. Mihailescu, E. A. Tanase, and P. Ordoñez de Pablos, "Educational recommender systems and their application in lifelong learning," *Behaviour & information technology*, vol. 35, no. 4, pp. 290–297, 2016.

[147] D. N. Beede, T. A. Julian, D. Langdon, G. McKittrick, B. Khan, and M. E. Doms, "Women in stem: A gender gap to innovation," *Economics and Statistics Administration Issue Brief*, no. 04-11, 2011.

[148] L. Sweeney, "Discrimination in online ad delivery," *Communications of the ACM*, vol. 56, no. 5, pp. 44–54, 2013.

[149] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[150] R. Marshall, "The economics of racial discrimination: A survey," *Journal of Economic Literature*, vol. 12, no. 3, pp. 849–871, 1974.

[151] S. L. Willborn, "The disparate impact model of discrimination: Theory and limits," *Am. UL Rev.*, vol. 34, p. 799, 1984.

[152] A. Romei and S. Ruggieri, "A multidisciplinary survey on discrimination analysis," *The Knowledge Engineering Review*, vol. 29, no. 5, pp. 582–638, 2014.

[153] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[154] M. Mansoury, "Fairness-aware recommendation in multi-sided platforms," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021, pp. 1117–1118.

[155] R. Burke, "Multisided fairness for recommendation,"

arXiv preprint arXiv:1707.00093, 2017.

[156] A. Gharahighehi, C. Vens, and K. Pliakos, "Fair multi-stakeholder news recommender system with hypergraph ranking," *Information Processing & Management*, vol. 58, no. 5, p. 102663, 2021.

[157] H. Wu, B. Mitra, C. Ma, F. Diaz, and X. Liu, "Joint multisided exposure fairness for recommendation," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.

[158] S. Yao and B. Huang, "Beyond parity: Fairness objectives for collaborative filtering," *Advances in neural information processing systems*, vol. 30, 2017.

[159] C. Wu, F. Wu, T. Qi, and Y. Huang, "Are big recommendation models fair to cold users?" *arXiv preprint arXiv:2202.13607*, 2022.

[160] Z. Fu, Y. Xian, R. Gao, J. Zhao, Q. Huang, Y. Ge, S. Xu, S. Geng, C. Shah, Y. Zhang *et al.*, "Fairness-aware explainable recommendation over knowledge graphs," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 69–78.

[161] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Correcting popularity bias by enhancing recommendation neutrality." in *RecSys*, 2014.

[162] H. Abdollahpouri, R. Burke, and B. Mobasher, "Controlling popularity bias in learning-to-rank recommendation," in *Proceedings of the eleventh ACM conference on recommender systems*, 2017, pp. 42–46.

[163] A. Ferraro, "Music cold-start and long-tail recommendation: bias in deep representations," in *Proceedings of the 13th ACM Conference on Recommender Systems*, 2019, pp. 586–590.

[164] H. Abdollahpouri, M. Mansoury, R. Burke, and B. Mobasher, "The unfairness of popularity bias in recommendation," *Workshop on recommendation in multi-stakeholder environments*, 2019.

[165] H. Abdollahpouri, R. Burke, and B. Mobasher, "Managing popularity bias in recommender systems with personalized re-ranking," in *The thirty-second international flairs conference*, 2019.

[166] L. Boratto, G. Fenu, and M. Marras, "Interplay between upsampling and regularization for provider fairness in recommender systems," *User Modeling and User-Adapted Interaction*, vol. 31, no. 3, pp. 421–455, 2021.

[167] F. Fabbri, F. Bonchi, L. Boratto, and C. Castillo, "The effect of homophily on disparate visibility of minorities in people recommender systems," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 165–175.

[168] M. D. Ekstrand, M. Tian, M. R. I. Kazi, H. Mehrpouyan, and D. Kluver, "Exploring author gender in book rating and recommendation," in *Proceedings of the 12th ACM conference on recommender systems*, 2018, pp. 242–250.

[169] E. Gómez, C. Shui Zhang, L. Boratto, M. Salamó, and M. Marras, "The winner takes it all: geographic imbalance and provider (un) fairness in educational recommender systems," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1808–1812.

[170] E. Gómez, L. Boratto, and M. Salamó, "Disparate impact in item recommendation: A case of geographic imbalance," in *European Conference on Information Retrieval*. Springer, 2021, pp. 190–206.

[171] ——, "Provider fairness across continents in collaborative recommender systems," *Information Processing & Management*, vol. 59, no. 1, p. 102719, 2022.

[172] Y. Wu, J. Cao, G. Xu, and Y. Tan, "Tfrom: A two-sided fairness-aware recommendation model for both customers and providers," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1013–1022.

[173] A. Chakraborty, A. Hannak, A. J. Biega, and K. Gummadi, "Fair sharing for sharing economy platforms," in *Fairness, Accountability and Transparency in Recommender Systems-Workshop on Responsible Recommendation*, 2017.

[174] G. K. Patro, A. Biswas, N. Ganguly, K. P. Gummadi, and A. Chakraborty, "Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms," in *Proceedings of The Web Conference 2020*, 2020, pp. 1194–1204.

[175] T. Sühr, A. J. Biega, M. Zehlike, K. P. Gummadi, and A. Chakraborty, "Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 3082–3092.

[176] L. Wang and T. Joachims, "User fairness, item fairness, and diversity for rankings in two-sided markets," in *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, 2021, pp. 23–41.

[177] E. Creager, D. Madras, T. Pitassi, and R. Zemel, "Causal modeling for fairness in dynamical systems," in *International Conference on Machine Learning*. PMLR, 2020, pp. 2185–2195.

[178] J. Williams and J. Z. Kolter, "Dynamic modeling and equilibria in fair decision making," *arXiv preprint arXiv:1911.06837*, 2019.

[179] X. Zhang, R. Tu, Y. Liu, M. Liu, H. Kjellstrom, K. Zhang, and C. Zhang, "How do fair decisions fare in long-term qualification?" *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 457–18 469, 2020.

[180] Y. Wu, R. Xie, Y. Zhu, F. Zhuang, A. Xiang, X. Zhang, L. Lin, and Q. He, "Selective fairness in recommendation via prompts," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 2657–2662.

[181] A. Bose and W. Hamilton, "Compositional fairness constraints for graph embeddings," in *International Conference on Machine Learning*. PMLR, 2019, pp. 715–724.

[182] S. Liu, Y. Ge, S. Xu, Y. Zhang, and A. Marian, "Fairness-aware federated matrix factorization," in *Proceedings of the 16th ACM Conference on Recommender Systems*, 2022, pp. 168–178.

[183] K. Maeng, H. Lu, L. Melis, J. Nguyen, M. Rabbat, and C.-J. Wu, "Towards fair federated recommenda-

tion learning: Characterizing the inter-dependence of system and data heterogeneity," in *Proceedings of the 16th ACM Conference on Recommender Systems*, 2022.

[184] S. Caton and C. Haas, "Fairness in machine learning: A survey," *arXiv preprint arXiv:2010.04053*, 2020.

[185] M. D. Ekstrand, M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, and M. S. Pera, "All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 172–186.

[186] R. Gao and C. Shah, "Addressing bias and fairness in search systems," in *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 2021, pp. 2643–2646.

[187] A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi *et al.*, "Fairness in recommendation ranking through pairwise comparisons," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2212–2220.

[188] R. Burke, N. Sonboli, M. Mansoury, and A. Ordoñez-Gauger, "Balanced neighborhoods for fairness-aware collaborative recommendation," 2017.

[189] G. Farnadi, P. Kouki, S. K. Thompson, S. Srinivasan, and L. Getoor, "A fairness-aware hybrid recommender system," *arXiv preprint arXiv:1809.09030*, 2018.

[190] L. Xiao, Z. Min, Z. Yongfeng, G. Zhaoquan, L. Yiqun, and M. Shaoping, "Fairness-aware group recommendation with pareto-efficiency," in *Proceedings of the Eleventh ACM Conference on Recommender Systems*, 2017, pp. 107–115.

[191] S. Yao and B. Huang, "New fairness metrics for recommendation that embrace differences," *arXiv preprint arXiv:1706.09838*, 2017.

[192] Y. Ge, X. Zhao, L. Yu, S. Paul, D. Hu, C.-C. Hsieh, and Y. Zhang, "Toward pareto efficient fairness-utility trade-off in recommendation through reinforcement learning," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022, pp. 316–324.

[193] A. Singh and T. Joachims, "Fairness of exposure in rankings," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2219–2228.

[194] T. Yang and Q. Ai, "Maximizing marginal fairness for dynamic learning to rank," in *Proceedings of the Web Conference 2021*, 2021, pp. 137–145.

[195] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates, "Fa* ir: A fair top-k ranking algorithm," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 1569–1578.

[196] L. E. Celis, S. Kapoor, F. Salehi, and N. Vishnoi, "Controlling polarization in personalization: An algorithmic framework," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 160–169.

[197] M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "Fair algorithms for infinite and contextual bandits," *arXiv preprint arXiv:1610.09559*, 2016.

[198] Y. Chen, A. Cuellar, H. Luo, J. Modi, H. Nemlekar, and S. Nikolaidis, "Fair contextual multi-armed bandits: Theory and experiments," in *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 181–190.

[199] S. C. Geyik, S. Ambler, and K. Kenthapadi, "Fairness-aware ranking in search & recommendation systems with application to linkedin talent search," in *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, 2019, pp. 2221–2231.

[200] A. Khademi, S. Lee, D. Foley, and V. Honavar, "Fairness in algorithmic decision making: An excursion through the lens of causality," in *The World Wide Web Conference*, 2019, pp. 2907–2914.

[201] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[202] C. Barabas, M. Virza, K. Dinakar, J. Ito, and J. Zittrain, "Interventions over predictions: Reframing the ethical debate for actuarial risk assessment," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 62–76.

[203] J. Zhang and E. Bareinboim, "Equality of opportunity in classification: A causal approach," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[204] N. Kilbertus, M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," *Advances in neural information processing systems*, vol. 30, 2017.

[205] J. Pearl, "Direct and indirect effects," in *Probabilistic and Causal Inference: The Works of Judea Pearl*, 2022, pp. 373–392.

[206] Y. Saito, "Asymmetric tri-training for debiasing missing-not-at-random explicit feedback," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

[207] X. Wang, R. Zhang, Y. Sun, and J. Qi, "Doubly robust joint learning for recommendation on data missing not at random," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6638–6647.

[208] J. Ma, R. Guo, M. Wan, L. Yang, A. Zhang, and J. Li, "Learning fair node representations with graph counterfactual fairness," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022, pp. 695–703.

[209] R. Mehrotra, J. McInerney, H. Bouchard, M. Lalmas, and F. Diaz, "Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems," in *Proceedings of the 27th acm international conference on information and knowledge management*, 2018, pp. 2243–2251.

[210] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2376–2384.

[211] L. Zhang, Y. Wu, and X. Wu, "A causal framework for discovering and removing direct and indirect discrimination," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, ser. IJCAI'17, 2017.

[212] ——, "Causal modeling-based discrimination discov-

ery and removal: criteria, bounds, and algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 11, pp. 2035–2050, 2018.

[213] ——, "On discrimination discovery using causal networks," in *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 2016, pp. 83–93.

[214] ——, "Situation testing-based discrimination discovery: A causal inference approach." in *IJCAI*, vol. 16, 2016, pp. 2718–2724.

[215] L. Zhang and X. Wu, "Anti-discrimination learning: a causal modeling-based framework," *International Journal of Data Science and Analytics*, vol. 4, no. 1, pp. 1–16, 2017.

[216] W. Huang, L. Zhang, and X. Wu, "Achieving counterfactual fairness for causal bandit," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 6, 2022, pp. 6952–6959.

[217] M. D. Ekstrand, J. T. Riedl, J. A. Konstan *et al.*, "Collaborative filtering recommender systems," *Foundations and Trends® in Human–Computer Interaction*, vol. 4, no. 2, pp. 81–173, 2011.

[218] Z. Ovaisi, S. Heinecke, J. Li, Y. Zhang, E. Zheleva, and C. Xiong, "Rgrecsys: A toolkit for robustness evaluation of recommender systems," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022, pp. 1597–1600.

[219] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, "Towards out-of-distribution generalization: A survey," *arXiv preprint arXiv:2108.13624*, 2021.

[220] L. Cao, "Non-iid recommender systems: A review and framework of recommendation paradigm shifting," *Engineering*, vol. 2, no. 2, pp. 212–224, 2016.

[221] Z. Wang, J. Zhang, H. Xu, X. Chen, Y. Zhang, W. X. Zhao, and J.-R. Wen, "Counterfactual data-augmented sequential recommendation," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 347–356.

[222] W. Wang, X. Lin, F. Feng, X. He, M. Lin, and T.-S. Chua, "Causal representation learning for out-of-distribution recommendation," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3562–3571.

[223] Z. Shen, P. Cui, K. Kuang, B. Li, and P. Chen, "Causally regularized learning with agnostic data selection bias," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 411–419.

[224] K. Kuang, P. Cui, S. Athey, R. Xiong, and B. Li, "Stable prediction across unknown environments," in *proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 1617–1626.

[225] P. Gutierrez and J.-Y. Gérardy, "Causal inference and uplift modelling: A review of the literature," in *International conference on predictive applications and APIs*. PMLR, 2017, pp. 1–13.

[226] N. Radcliffe, "Using control groups to target on predicted lift: Building and assessing uplift model," *Direct Marketing Analytics Journal*, pp. 14–21, 2007.

[227] M. Jaskowski and S. Jaroszewicz, "Uplift modeling for clinical trial data," in *ICML Workshop on Clinical Data Analysis*, vol. 46, 2012, pp. 79–95.

[228] N. J. Radcliffe and P. D. Surry, "Real-world uplift modelling with significance-based uplift trees," *White Paper TR-2011-1, Stochastic Solutions*, pp. 1–33, 2011.

[229] P. Rzepakowski and S. Jaroszewicz, "Decision trees for uplift modeling with single and multiple treatments," *Knowledge and Information Systems*, vol. 32, no. 2, pp. 303–327, 2012.

[230] M. Sato, H. Izumo, and T. Sonoda, "Modeling individual users' responsiveness to maximize recommendation impact," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, 2016, pp. 259–267.

[231] A. V. Bodapati, "Recommendation systems with purchase data," *Journal of marketing research*, vol. 45, no. 1, pp. 77–93, 2008.

[232] E. A. Stuart, "Matching methods for causal inference: A review and a look forward," *Statistical science: a review journal of the Institute of Mathematical Statistics*, vol. 25, no. 1, p. 1, 2010.

[233] M. Sato, S. Takemori, J. Singh, and T. Ohkuma, "Unbiased learning for the causal effect of recommendation," in *Fourteenth ACM Conference on Recommender Systems*, 2020, pp. 378–387.

[234] A. Swaminathan and T. Joachims, "The self-normalized estimator for counterfactual learning," *advances in neural information processing systems*, vol. 28, 2015.

[235] T. Xiao and S. Wang, "Towards unbiased and robust causal ranking for recommender systems," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022, pp. 1158–1167.

[236] M. J. Funk, D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian, "Doubly robust estimation of causal effects," *American journal of epidemiology*, vol. 173, no. 7, pp. 761–767, 2011.

[237] A. J. Chaney, B. M. Stewart, and B. E. Engelhardt, "How algorithmic confounding in recommendation systems increases homogeneity and decreases utility," in *Proceedings of the 12th ACM conference on recommender systems*, 2018, pp. 224–232.

[238] D. Jannach, L. Lerche, I. Kamehkhosh, and M. Jugovac, "What recommenders recommend: an analysis of recommendation biases and possible countermeasures," *User Modeling and User-Adapted Interaction*, vol. 25, no. 5, pp. 427–491, 2015.

[239] M. Mansoury, H. Abdollahpouri, M. Pechenizkiy, B. Mobasher, and 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.

[240] Y. Ge, S. Zhao, H. Zhou, C. Pei, F. Sun, W. Ou, and Y. Zhang, "Understanding echo chambers in e-commerce recommender systems," in *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, 2020, pp. 2261–2270.

[241] D. P. Allen, H. J. Wheeler-Mackta, and J. R. Campo, "The effects of music recommendation engines on the filter bubble phenomenon," *Interactive Qualifying Projects*, 2017.

[242] U. Chitra and C. Musco, "Analyzing the impact of

filter bubbles on social network polarization," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, pp. 115–123.

[243] C. Gao, W. Lei, J. Chen, S. Wang, X. He, S. Li, B. Li, Y. Zhang, and P. Jiang, "Cirs: Bursting filter bubbles by counterfactual interactive recommender system," *arXiv preprint arXiv:2204.01266*, 2022.

[244] T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan, "Exploring the filter bubble: the effect of using recommender systems on content diversity," in *Proceedings of the 23rd international conference on World wide web*, 2014, pp. 677–686.

[245] J. McInerney, B. Brost, P. Chandar, R. Mehrotra, and 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.

[246] A. Agarwal, K. Takatsu, I. Zaitsev, and T. Joachims, "A general framework for counterfactual learning-to-rank," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 5–14.

[247] A. Agarwal, I. Zaitsev, X. Wang, C. Li, M. Najork, and T. Joachims, "Estimating position bias without intrusive interventions," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 474–482.

[248] Q. Ai, K. Bi, C. Luo, J. Guo, and W. B. Croft, "Unbiased learning to rank with unbiased propensity estimation," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 385–394.

[249] M. Chen, C. Liu, J. Sun, and S. C. Hoi, "Adapting interactional observation embedding for counterfactual learning to rank," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 285–294.

[250] T. Joachims, A. Swaminathan, and T. Schnabel, "Unbiased learning-to-rank with biased feedback," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2017, pp. 781–789.

[251] Z. Qin, S. J. Chen, D. Metzler, Y. Noh, J. Qin, and X. Wang, "Attribute-based propensity for unbiased learning in recommender systems: Algorithm and case studies," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2359–2367.

[252] J.-w. Lee, S. Park, and J. Lee, "Dual unbiased recommender learning for implicit feedback," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1647–1651.

[253] Z. Zhao, J. Chen, S. Zhou, X. He, X. Cao, F. Zhang, and W. Wu, "Popularity bias is not always evil: Disentangling benign and harmful bias for recommendation," *arXiv preprint arXiv:2109.07946*, 2021.

[254] P. Brouillard, S. Lachapelle, A. Lacoste, S. Lacoste-Julien, and A. Drouin, "Differentiable causal discovery from interventional data," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 865–21 877, 2020.