A Survey on Trustworthy Recommender Systems

YINGQIANG GE, Department of Computer Science, Rutgers University, USA
SHUCHANG LIU, Department of Computer Science, Rutgers University, USA
ZUOHUI FU, Department of Computer Science, Rutgers University, USA
JUNTAO TAN, Department of Computer Science, Rutgers University, USA
ZELONG LI, Department of Computer Science, Rutgers University, USA
SHUYUAN XU, Department of Computer Science, Rutgers University, USA
YUNQI LI, Department of Computer Science, Rutgers University, USA
YIKUN XIAN, Department of Computer Science, Rutgers University, USA
YONGFENG ZHANG, Department of Computer Science, Rutgers University, USA

Recommender systems (RS), serving at the forefront of Human-centered AI, are widely deployed in almost every corner of the web and facilitate the human decision-making process. However, despite their enormous capabilities and potential, RS may also lead to undesired counter-effects on users, items, producers, platforms, or even the society at large, such as compromised user trust due to non-transparency, unfair treatment of different consumers, or producers, privacy concerns due to extensive use of user's private data for personalization, just to name a few. All of these create an urgent need for *Trustworthy Recommender Systems (TRS)* so as to mitigate or avoid such adverse impacts and risks. In this survey, we will introduce techniques related to trustworthy and responsible recommendation, including but not limited to explainable recommendation, fairness in recommendation, privacy-aware recommendation, robustness in recommendation, user controllable recommendation, as well as the relationship between these different perspectives in terms of trustworthy and responsible recommendation. Through this survey, we hope to deliver readers with a comprehensive view of the research area and raise attention to the community about the importance, existing research achievements, and future research directions on trustworthy recommendation.

1 INTRODUCTION

Recommender systems (RS)—which are extensively deployed in various systems such as e-commerce, social networks, search engines, news portals, hiring platforms, intelligent assistants, smart home and smart city services, as well as healthcare and financial applications—have been acknowledged for their capacity to deliver high-quality services that bridge the gap between users and items by delivering tailored content for each individual. recommender systems not only help users to find relevant information more efficiently, but also directly influence the human decision-making process by providing relevant suggestions or even shape users' worldviews by exposing users to the selected content. Overall, recommender system is the frontier of Human-centered AI research and works as the bridge between humans and AI.

However, for every plus there is a minus, RS may offer both promise and perils. There are growing concerns that the irresponsible use of recommendation techniques may bring counter-effects and untrustworthy issues, such as compromised user trust due to non-transparency, unfair treatment of different users, producers, or platforms, privacy concerns due to extensive use of user's private data for personalization, echo chambers due to the lack of controllability for users that leads to repeated

Authors' addresses: Yingqiang Ge, Department of Computer Science, Rutgers University, USA, yingqiang.ge@rutgers.edu; Shuchang Liu, Department of Computer Science, Rutgers University, USA, sl1471@scarletmail.rutgers.edu; Zuohui Fu, Department of Computer Science, Rutgers University, USA, zuofui.fu@rutgers.edu; Juntao Tan, Department of Computer Science, Rutgers University, USA, juntao.tan@rutgers.edu; Zelong Li, Department of Computer Science, Rutgers University, USA, zelong.li@rutgers.edu; Shuyuan Xu, Department of Computer Science, Rutgers University, USA, shuyuan.xu@rutgers.edu; Yunqi Li, Department of Computer Science, Rutgers University, USA, yunqi.li@rutgers.edu; Yikun Xian, Department of Computer Science, Rutgers University, USA, yx150@rutgers.edu; Yongfeng Zhang, Department of Computer Science, Rutgers University, USA, yongfeng.zhang@rutgers.edu.

reinforcement of users' existing interests — the list just continues to expand. These vulnerabilities significantly limit the development and deployment of recommendation algorithms and may even lead to severe economic and social issues. As a result, only considering recommendation accuracy is not enough when developing modern recommendation systems. We also need to make sure that the models are fair, have not been tampered with, will not fall apart in different conditions, and can be understood by humans. Moreover, the design and development process of RS also needs to be transparent and inclusive. All of these considerations beyond accuracy that makes recommender systems safe, responsible, and worthy of our trust are related to trustworthy recommender systems research. Since recommender system is an important direction of Human-centered AI research that directly involves humans in the loop, Trustworthy Recommender System (TRS) has been leading the research of Trustworthy Artificial Intelligence (TAI) over the past years on various perspectives, such as the definition, method and evaluation of trustworthiness on explainability, fairness, robustness, and privacy, as well as how humans interact with trustworthy AI systems.

Therefore, as a vital instantiation of trustworthy AI under the context of recommender system research, in this survey, we introduce trustworthy recommender systems as competent RS that incorporates the core aspects of trustworthiness such as *explainability, fairness, privacy, robustness* and *controllability*. We believe that incorporating these aspects when designing recommender systems will improve their responsibility, gain trust from human users, and significantly promote recommender systems for social good.

Differences from Existing Surveys in Recommendation. There have already been several recent surveys focusing on specific ethical topics in recommendation scenario, such as explainability [61, 340], bias and fairness [54, 77, 178, 225, 288, 328], privacy protection [140, 308], user controllability [142, 143], etc. These surveys successfully highlight the importance of the social responsibility of recommender systems, leading to further developments in this important line of research. However, these topics have only been presented in their own self-contained ways, while a systematic view of trustworthiness in recommendation and the inherent relationship among the various trustworthiness perspectives is highly needed. The closest work to ours are Dong et al. [82] and Mobasher et al. [207]. However, [82] only covers user social relationship, robustness, and explainability, while [207] only discusses the attack models and algorithm robustness in recommendation, and they did not examine the internal relationship among these concepts. In contrast, our work introduces trustworthiness over more comprehensive perspectives, highlights the relationship among the perspectives, and sheds light on open problems and future directions to explore the intersection of the perspectives.

Relationship with other Trustworthy AI Research. Due to its importance and necessity, there have been numerous discussions and debates over the implication of Trustworthy Artificial Intelligence (TAI). In particular, Toreini et al. [273] study trust in AI and summarized the attributes of TAI as Ability, Benevolence, Integrity, and Predictability; Varshney [280] believe that a trustworthy machine learning system is one that has sufficient Basic Performance, Reliability, Human Interaction, and Selflessness; Liu et al. [186] consider TAI as threat or risk-free programs and focus on six dimensions in achieving trustworthiness: Safety & Robustness, Nondiscrimination & Fairness, Explainability, Privacy, Accountability & Auditability, and Environmental Well-being. Moreover, in the year of 2019, European Union (EU) proposed the Ethics Guidelines for Trustworthy AI[1], requiring that an AI system should meet four ethical principles: respect for human autonomy, prevention of harm, explicability, and fairness [6]. Although the existing literature explores the space of trustworthiness from various perspectives, several key aspects that have received the most

[1]https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

recognition and consensus are *explainability, fairness, privacy, robustness, and controllability*, which we believe are also the key components of TRS.

Intended Audience and Paper Organization. The primary readers of the survey are RS researchers, technologists, and practitioners aiming to make recommender systems more trustworthy and responsible with their expertise. On the other hand, since recommender system is a very representative and pervasive human-centered AI system, thus the target readers of the survey also include general AI researchers, practitioners, theorists as well as public policymakers who are interested in trustworthiness, ethics, and regulations of AI. The rest of the survey is organized as follows: Section 2 introduces the preliminary knowledge of recommender systems and some representative recommendation algorithms. Section 3, 4, 5, 6, 7 focus on explainability, fairness, privacy, robustness, and controllability, respectively. The last section concludes the survey.

2 RECOMMENDER SYSTEM BASICS

Before presenting trustworthy recommendation in detail, we will briefly introduce the basic problems of personalized recommendation and provide a formalized representation.

2.1 The Input and Output of RS

Recommender systems may have various forms of possible input data. A basic recommender system usually has three types of input data: user, item, and interaction, where the interaction could be users' click, purchase, rating, review, or other behaviors over the items (Figure 1). We should notice that the "item" here has a very broad meaning and could be various types of things. For example, it could be products on an e-commerce shopping website, tweets or ads on social networks, hotels or air tickets on trip planning websites, videos or music in media streaming applications, jobs in online marketplaces, or even other users of this website such as in friend recommendation. Different types of items have different attributes. For example, products of shopping sites may have brands, manufacturers, size, and weight, while for media recommendation, the attributes may include the type of media, style, and content descriptions, and trip recommendation may focus on travel methods, time duration, stop numbers and estimated cost, etc. Besides, the "user" may not only be an ID in the recommender system, but also a profile that can describe the user. User profiles can have distinct forms in different application scenarios or in different recommendation algorithms. An intuitive and understandable form of the user profile is the registration information of the user, such as age, gender, annual income, location, and active time intervals.

Apart from users and items, interaction is an important input of recommender system as well, which is the connection between users and items. One of the most commonly used types of interaction is the rating information from users to items. The 5-star rating scale is widely used on major websites, which reflects the user's preference for the item. Rating is denoted as an integer between 1 and 5 (1 and 5 included) in most recommendation algorithms. The preference of users may also be revealed by other interactions, such as users' reviews, clicks, and purchase histories. Usually, we can separate these contents into two types, explicit feedback, and implicit feedback. Explicit feedback such as ratings and reviews are collected when users actively and explicitly tell the system about their preferences for an item, while implicit feedback such as user clicks are passively recorded when users interact with the website interface, and the data can implicitly reflect the users' preference. For instance, if a user explores a lot of detailed information about an item or the user spends a lot of time viewing the web page of the item, then it implies that the user may have a strong interest in the item.

The output of a recommender system usually includes a personalized recommendation list tailored to the user and the explanations accompanying the recommendations (Figure 2). The process of generating the output usually includes three stages: Predict, Rank, and Explain (PRE).

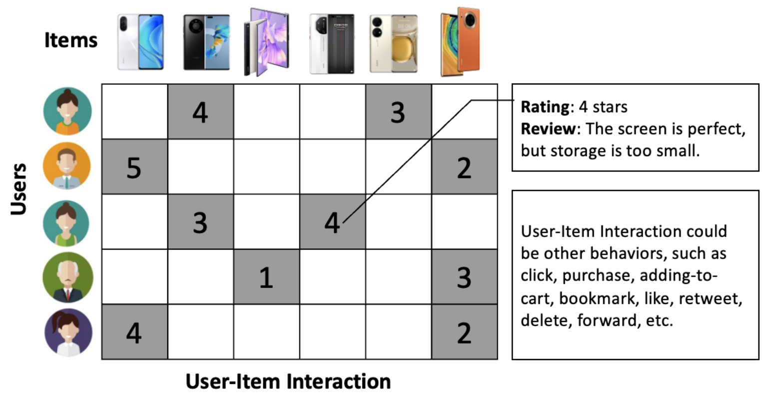

Fig. 1. Typical inputs of a recommender system.

The predict stage aims to make preliminary predictions about the user's preferences on the items and generate a preliminary list of candidate items for the user. The preliminary list could be very long such as thousands or more, as a result, this stage is sometimes also called the "recall stage" in industrial production environments. The ranking stage refines the list from the previous predict stage and generates the final recommendation list for the user, which usually includes just a few or a few tens of items. Although the easiest way of ranking is just sorting the items according to the predicted scores in the predicting stage, a lot of other important factors play a role in this stage, such as refining the preference predictions based on more information and more meticulously designed algorithms that can only be executed on the smaller amount of items selected by the previous stage, re-ranking the list based on diversity, fairness, serendipity, business or other considerations, and responding to user's real-time requests. Finally, the explain step generates explanations for each recommended item to justify the recommendation. Depending on the explainable recommendation method used to generate the explanations, the explain step may either happen simultaneously with the ranking step or after the ranking step: if the ranking model used in the ranking step is intrinsically explainable, then the model can directly generate explanations alongside the recommendation list, and thus the explain step happens at the same time; otherwise, if the ranking model is black-box, then a post-hoc explanation generation model is needed in the explain step to generate the explanations after the recommendation list is provided.

2.2 Representative Recommendation Algorithms

There are a diverse set of recommendation algorithms. In this survey, we broadly categorize the recommendation algorithms into three stages: shallow models, deep models, and big models.

- **Shallow Models**: The pioneer works of recommender systems started with shallow models which take expert-designed similarity functions to extract simple and effective user-item matching patterns from data. This includes both collaborative filtering approaches [156, 183, 240, 243] and content-based filtering approaches [17, 278]. Collaborative Filtering (CF) is one of the most fundamental approaches for recommender systems, which has been widely used in real-world systems. Early CF methods adopted simple similarity functions such as cosine similarity and inner product for matching. The basic idea is that similar users may share similar interests and similar items may be liked by similar users. CF methods can be further divided into two groups according to whether user-item similarity is calculated based on direct user/item features or learned user/item features. Direct-feature based CF is also called memory-based CF [36], which calculates user similarity or item similarity for recommendation based on a similarity measurement over

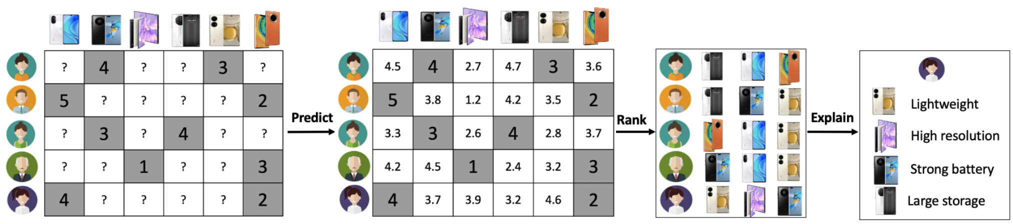

Fig. 2. Typical workflow and output of a recommender system.

users' and/or items' direct historical data. For example, User-based CF [156, 240] directly takes each row vector in the user-item rating matrix as the representation for each user, and calculates the user similarity based on Pearson correlation coefficient or cosine similarity; Item-based CF [183, 243] takes the column vector of the user-item rating matrix as the representation for each item, and also employs similarity functions to calculate the similarity between items for recommendation. Learned-feature based CF is also called model-based CF [36], which adopts a model to learn the user and item representations for making predictions. Such models are also frequently named Latent Factor Models (LFM). For example, Matrix Factorization (MF) [157] learns each user and item as a latent embedding vector and takes the inner product between the user and item vectors to calculate the user-item similarity for recommendation; Probabilistic Matrix Factorization (PMF) also learns user and item latent representations for recommendation and provides a probabilistic interpretation for matrix factorization under the Bayesian learning framework [206]. Content-based filtering approach, on the other hand, relies on the various user and item content features such as user profiles and item descriptions for recommendation [17, 30, 223, 278]. For example, a movie can be described by its features such as genre, director, cast, length, and language. Content-based filtering calculates the similarity between items based on their content similarity (e.g., same director, overlapping cast) and recommends those movies to a user that are contently similar to the user's previously watched movies [17, 155, 223, 278].

- **Deep Models**: The development of deep learning and neural networks have further improved recommendation methods [336]. The various neural approaches to recommendation can be broadly classified into two categories: Collaborative Filtering (CF) [7, 65, 127, 137, 196, 312, 339, 345] and Collaborative Reasoning (CR) [52, 53, 248, 324, 337, 338], while the CF approach can be further classified into the similarity learning methods [65, 127, 137, 312] and the representation learning methods [7, 196, 339, 345]. Deep learning based CF considers recommendation as a perceptive learning problem, which employs similarity learning or representation learning to extract perceptive correlation patterns from data for matching and recommendation; while deep learning based CR considers recommendation as a cognitive reasoning problem, which employs logical reasoning or causal reasoning for user behavior prediction and recommendation. More specifically, for deep learning based CF, the similarity learning approach adopts simple user/item representations (such as one-hot vector) and learns a complex matching function (such as a neural prediction network) to calculate user-item matching scores [65, 127, 137, 312], while the representation learning approach learns rich user/item representations from text, image, knowledge, etc. and adopts a simple matching function (e.g., inner product) for efficient matching score calculation [7, 196, 339, 345]. It is worth noting that there exist debates over whether complex matching functions are better than simple functions [73, 74, 91, 238]. The above recommendation algorithms use a user's historical interaction to learn the static preference, however, in many real-world scenarios, the next behavior of a user not only depends on the static long-term preference but also relies on the current intent [88]. Therefore, sequential recommendation (also related to session-based or session-aware recommendation), which models the user behavior as a sequence for future behavior prediction, has become increasingly important in academia and industry [59, 75, 131, 139, 149, 188, 193, 259, 268, 346]. Traditional sequential recommendation models employ simple machine learning approaches to model sequential data, such as Markov Chain [237] and Session-based KNN [138]. With the development of deep learning techniques, many deep models obtain tremendous achievements in sequential recommendation, including RNN [131], CNN [268, 321], LSTM [296], BERT [75, 259], attention models [149, 284], and memory networks [59]. Another important direction is learning to rank for recommendation [3, 125, 236], which learns the relative ordering of items instead of the absolute preference scores. A representative method for learning to rank in recommendation is Bayesian personalized ranking (BPR) [236], which is a pair-wise learning to rank method. It is also further generalized to take other information sources such as images [125].

- **Big Models**: Recently, Foundation Models such as Large Language Models (LLM) have achieved surprisingly good performance in many AI sub-fields, which have the advantages of emergent capabilities from model size, extracting useful information based on self-supervision, unifying various downstream tasks based on pre-training, fine-tuning and prompting, as well as generalizing to zero-shot or few-shot cases [31, 290]. Many powerful foundation models have been developed for natural language tasks such as T5 [231], GPT-3 [35], OPT [335] and PaLM [69], which show impressive performance on language understanding, generation and reasoning tasks. Since language grounding is a powerful medium that can describe almost any data, problem, or task, many other problems can be formulated as language sequences by learning the corresponding token embeddings and integrating them with normal word embeddings. For example, many pre-trained vision-language models based on visual token embedding have demonstrated strong power in visual understanding, generation, and reasoning tasks as well as vision-language co-learning tasks, such as CLIP [230] and DALL-E [232, 233]. Recommender system research has demonstrated a similar trend, besides, since personalization is one of the most unique and important characteristics of recommender system research, the recommender system community has been leading the research on Personalized Foundation Models (PFM). One prominent example is P5 [108], which is a Pretrain, Personalized Prompt, and Predict Paradigm for recommendation. P5 reformulates recommendation as a language understanding and generation task based on personalized prompts, and it unifies various recommendation tasks such as rating prediction, direct recommendation, sequential recommendation, explanation generation, and review summarization, which leads to a universal recommendation engine. As a result, P5 serves as the foundation model for various downstream recommendation tasks.

In the following sections, we will discuss how to encode various trustworthiness considerations into recommendation models and recommender system research. In particular, we will discuss explainability, fairness, privacy, robustness, controllability, and other perspectives to consider in trustworthy recommender systems research.

3 EXPLAINABILITY

Explainable recommendation has been an important area in both industry and academia, aiming to improve transparency, user satisfaction, and trustworthiness over the recommender systems [340, 341]. Specifically, the goal is to provide comprehensible justifications along with recommended items to help stakeholders make better and solid decisions and simultaneously improve the transparency and trustworthiness of recommender systems. Explanation in recommender systems can be leveraged to assist model developers to understand and debug the working mechanism of the decision-making process, and also to facilitate better engagement and trustworthiness for the end users who consume the results produced by the systems. In addition to web portal based recommender systems, explanation is also integrated into conversational recommendation interfaces such as Apple Siri, Microsoft Cortana, and Amazon Alexa that serve as the direct interactive portal for end-users [62]. They are able to provide clear elicitation of user preference, interact with users intelligently and offer explanations with certain suggestions.

3.1 Overview of Explainable Recommendation

The term "Explainable Recommendation" was first defined by Zhang et al. [341]. As an important sub-field of AI and machine learning research and due to the fact that recommendation naturally involves humans in the loop, the recommender system community has been leading the research on Explainable AI ever since, which triggers a broader scope of explainability research in other AI and machine learning sub-fields [71, 340], such as explainability in scientific research [181], computer vision [297], natural language processing [40, 106, 172, 217, 229], graph neural networks [265, 299], database [112, 291], healthcare systems [121, 228, 350], online education [9, 20, 216, 264, 277], psychological studies [271] and cyber-physical systems [10, 12, 134, 135, 241].

Explainable recommendations provide additional explanations on the predicted results to better understand the inference and reasoning process behind black-box prediction models. As a crucial part of the modern paradigm AI, deep neural networks have increasingly contributed to the development of most state-of-the-art machine learning systems. However, these deep neural models are treated as black-box since they are too complicated and opaque to understand [83, 102, 307, 336]. Most of them are not designed toward explainability and transparency, which leads to negative consequences in many human-centered scenarios [119, 208]. For instance, if a medical-diagnosis system fails to provide supportive evidence to justify why the prediction is correct or not, the doctor can hardly adopt the automated decision even if its actual accuracy is high. Same situation can be applied to other domains such as e-commerce [304] and digital marketing [303].

The fundamental goal of explainability in AI is to open up the "black-box" of the pipelines in AI-related domain, not only to provide trustworthy explanations to users, but also to drive toward more interpretable models. As an important type of intelligent decision-making system, the modern recommender system is expected to provide high-quality recommendation results as well as personalized and intuitive explanations with better user engagement, which are important for many practical applications such as e-commerce and social media platforms. In addition, recommender systems manifest special characteristics on the requirement of explainability in the following aspects.

- **Personalized Explanation**. Some existing works claim that recommender systems should provide different recommendations and explanations to fit different user preferences. Li et al. [172] explore the personalized natural language generation in the review summarization and dialog systems. The proposed method addresses personalized generation based on Transformer architecture so as to bootstrap the strength of language modeling to generate high-quality and personalized explanations for recommender systems. Chen et al. [57] propose a novel neural architecture for explainable recommendation over fashion clothes category based on both image region-level features and review from user information. The corresponding personalized explanation will be highlighted through some regions of the image shown to the users. Li et al. [173] study user-understandable explanations using prompt learning. By sequential tuning and regarding recommendation process as regularization, it successfully fuses item IDs into the models to generate natural language explanations for recommendations by treating user and item IDs as prompts.

- **Interactive Feedback**. Providing explanations to users may have downstream impacts, especially in the applications of session recommendation, conversational recommendation, and interactive recommendations, compared to general ML tasks where explanations are only associated with one-time predictions. For example, Chen et al. [62] introduce an explainable recommender system during conversations between the users and agents. High explanation quality is guaranteed through multi-turn user model conversation. Omidvar-Tehrani et al. [215] conduct Explainable Points-of-Interest (POI) Recommendation as an exploratory process in which users are allowed to keep interaction with the system explanations by expressing their favorite POIs, and the interactions will impact the recommendation process. Wu et al. [297] provide vision-language explanations by deconfounded learning that conducts pre-training for the vision-language model. In that way, the potential effects of confounders are removed, which will expedite accurate representation training and better explainability. The proposed interactive mechanisms help users to better understand why the system returns particular results and also allow users to effectively return feedback to improve the recommender results.
- **Subjective Reaction**. General interpretable machine learning methods populate explanations for understanding the underlying mechanism of how models make predictions so that model developers can better tune or debug the methods. In contrast, explainable recommendations are devised for end users as well, who may have no knowledge about AI at all and are more subjective to understand the explanations. As a result, users may react to the explanations differently. Chen et al. [61] claim that recommender system should be considered as a subjective AI task where there is no definitely "right" or "wrong" item to recommend for a user. Instead, it all depends on whether the system can appropriately explain and justify the recommendation. Among the methods and evaluation metrics, the satisfaction and trust of users should be carefully considered. The framework in Le and Lauw [166] incorporates both subjective and objective aspect-level quality assumptions and integrates them with recommendation objectives as constraints.

3.2 Taxonomy on Explainable Recommendation Approaches

Considering the specialty and importance of explainable recommendations, many methods have been proposed and researched. Current research on explainable recommendations generally considers the following orthogonal perspectives.

- **Explanation Method**: Model-intrinsic vs Model-agnostic. It refers to whether the explanation generation process is part of the recommendation step inside the recommendation model or a post-hoc step that generates explanations after the recommendation model has provided the recommended items.
- **Explanation Scope**: Global vs Local explanation. It means whether the explanation is for the general recommendation model as a whole or for an individual recommended item.
- **Explanation Style**: The inherited form of explanations usually depends on the input such as entities (users or items), interactions, text, images, graphs, and others. Inherited form refers to the explanation directly generated by the model without post-processing or rendering steps, since any inherited forms of explanations can be converted into other formats such as textual narratives.
- **Benefited Users**: For the customer, explainable recommender system will improve user experience in the recommendation process and help them better understand recommended items with convincing justifications. On the other hand, explainable recommender system is able to help model developers debug black-box models and dissect the mechanism behind the model so as to improve RS in future iterations. It also helps content providers to understand whether the content is proper for a group of targeted users. Last but not least, explainable recommendations

will provide evidence to auditors so as to build trustworthiness among users, platforms, and third-party authorities under regulations such as GDPR [2].

In this survey, we categorize existing works on explainable recommendation based on input data type, since it directly determines different types of methods and inherited forms of output explanations. In parallel, for each subset of methods, we also label them with explanation method and explanation scope so as to make it easier for readers to understand the different dimensions of the taxonomy.

Entities. Entity-based explanation refers to a single instance such as similar users, items, or attributes, which can be used as an explanation for the recommended items. For instance, Explicit Factor Model (EFM) [341, 343] generates global explainable recommendations via extracting explicit item features as well as user opinions from user reviews in a model-agnostic way. Tree-enhanced embedding model (TEM) [286] learns the decision rules at the first stage, and then an embedding model is designed to incorporate the cross features and generalize the hidden user and item IDs. Dynamic Explainable Recommender (DER) [60] proposes the explainable recommendation paradigm dynamically based on the time-aware gated recurrent unit (GRU) modules and represents items based on sentence-level convolutional neural network (CNN). Extract-Expect-Explain (EX^3) [304] generates a set of relevant and similar items as recommendations in the e-commerce area, and a corresponding set of similar attributes is selected to justify the recommended same group of items. Neural Collaborative Reasoning (NCR) and related works [52, 53, 248, 300, 337, 349] employ explicit neural-symbolic reasoning rules over users, items or attributes to make the recommendation process transparent.

Text. Textual data widely exists in recommender systems such as item descriptions and user reviews. Some explainable recommendation algorithms extract important information from the input text and yield understandable auxiliary sentence justifications as explanations. For instance, Wang et al. [285] develop a model-agnostic reinforcement learning framework to generate textual sentence explanations. Chen et al. [49] adopt sequence-to-sequence modeling to generate textual sentence explanations. Li et al. [172, 173] propose personalized transformers and personalized prompt learning to generate fluent and high-coverage explanation sentences, which show significant explanation generation quality. Pan et al. [220] utilize the review rationalization with rationale generator to extract rationales from reviews to alleviate the effects of spurious correlations when explaining rating predictions. The experimental results demonstrate the improvement of recommendation accuracy as well as novel causal-aware explanations. Hada and Shevade [120] present an end-to-end framework for explaining recommendations with a sentiment classifier to control the pre-trained language model. The benefit is bypassing the costly training from scratch of the language model which is efficient for generating reviews. Wang et al. [283] develop a multi-task learning framework for explainable recommender system. Specifically, both user preference and content modeling are jointly learned by tensor factorization.

Multimedia. This type of data refers to rich multimedia such as images and video, which can provide more intuitive and fascinating demonstrations of items. Cheng et al. [66] apply an aspect-aware topic model as the multi-modal over both text reviews and item images so as to better model user preferences and item features from different aspects. The estimation of the aspect importance is also integrated into an aspect-aware latent factor model. The proposed framework alleviates the data sparsity challenge and presents good explainability for recommendation.

[2] https://en.wikipedia.org/wiki/General_Data_Protection_Regulation

Logical and Neural-Symbolic Rules. Neural-Symbolic rule-based recommender systems can conduct reasoning based on predefined or learned logical rules to make predictions, and the rules can be used to explain the recommendation process. Shi et al. [248] unify the power of deep learning and logic reasoning and propose a dynamic neural-symbolic architecture which enables logical reasoning in a differentiable representation space. Some basic logical operations such as AND, OR, NOT are learned as neural modules based on self-supervised logical regularizer to infer to true or false value of logical expressions. Chen et al. [53] further propose Neural Collaborative Reasoning (NCR), which takes explicit reasoning rules for improved transparency in recommendation. Zhu et al. [349] and Xian et al. [302] propose knowledge graph (KG) enhanced neural-symbolic reasoning model for recommendation by marrying the interpretability of symbolic rules and the expressiveness of KG embeddings. Neural-symbolic reasoning is able to guide the path reasoning process to generate faithful explanations, which are demonstrated to be consistent with historic user behaviors and the resulting paths genuinely reflect the decision-making process in KG reasoning. Recently, Zhang et al. [337] propose attribute-level neural-symbolic reasoning for recommendation, which extracts attribute-level logical rules to further enhance the transparency of the recommendation process.

Graphs. Graph is a widely-used data format in recommendation such as user-item bipartite graph, item-attribute knowledge graph, and user-interest temporal graph. As a result, mining explanations from the graph structure is important. Wang et al. [287] explore path representations by composing the semantics of both entities and relations within the KG. The reasoning procedure from user entities toward item entities over the KG paths will highlight the explanation process of the recommendation. The pivotal work Xian et al. [301] propose a reinforcement knowledge graph reasoning framework, which is able to generate path-based explanations via a policy-guided path reasoning (PGPR) agent to conduct efficient multi-hop reasoning over knowledge graphs. The resulting reasoning paths serve as the explanation since they explicitly expose the multi-step decision-making procedure. Xian et al. [300, 302] further propose neural-symbolic reasoning over knowledge graphs for explainable recommendation, which learns graph relations as neural-symbolic operators and encodes KG paths as user profiles based on neural-symbolic computation to capture prominent user behaviors in graph reasoning. Geng et al. [107] consider KG paths as sequences of tokens and proposed Path Language Modeling (PLM) for explainable recommendation. Such method includes learning a language model over the knowledge-grounded KG paths for path sequence decoding. Moreover, the method is capable of conducting explainable recommendation even when the KG structure is sparse and extremely large. Ma et al. [194] propose a joint learning framework which utilizes knowledge graphs to induce associative explainable rules for item through rule learning. By extracting rules, the method also shows a better ability to deal with the cold-start recommendation problem.

Counterfactuals. Some existing works adopted counterfactual reasoning from causal inference and applied it in recommendation scenarios. For a RS model that takes the input data (e.g., user history, item features.) and makes certain recommendations, counterfactual reasoning looks for what input should be changed and, by how much, to acquire a different prediction. The changed factors could be essential to form the explanation. More specifically, Ghazimatin et al. [111] propose a searching algorithm on heterogeneous graph, which looks for a minimal set of users' historical actions such that by removing them, the model will recommend different items. Xu et al. [310] propose a causality mining algorithm based on input sequence perturbation to extract counterfactual explanations for sequential recommendation. Tan et al. [266] propose a general explainable recommendation framework, Counterfactual Explainable Recommendation (CountER). It formulated a machine learning optimization problem to generate simple and effective explanations based on item aspects. Moreover, [266] also propose a causal evaluation metric to quantitatively evaluated the faithfulness of the generated explanations without access to the ground truth data. Later, as an extension of [266], [265] further analyze the relationship between counterfactual and factual reasoning and suggested that by combing both counterfactual and factual reasoning into one framework, the explainable models could generate explanations that are both sufficient and necessary. Tran et al. [275] extend influence functions to identify the most relevant training points and deduce a counterfactual set as explanation.

Multi-round interactions. Chen et al. [62] introduce explainable conversational recommendation. In the multi-turn user modeling conversational recommender system, a novel multi-task learning framework that enables tight collaboration between recommendation prediction, explanation generation, as well as user feedback integration is formulated. The multi-view feedback integration process integrates user feedback into the recommendation explanations so as to help users understand the model and meanwhile collect user feedback to understand the user interests. Fu et al. [97] propose a Human-in-the-Loop (HitL) graph reasoning paradigm and created a benchmark dataset to support explainable conversational recommendation research over knowledge graphs. Specifically, the idea is to leverage knowledge graphs to interpret diverse user behaviors. The conversational turns are able to track the human decision-making process while tracing the knowledge graph structures for transparency and explanation generation.

3.3 Evaluation of Explanations

A fundamental problem of explainable recommendation is how to evaluate the explanations [61, 340]. According to previous works [272], the explanation perspectives can be categorized based on four groups of serving targets:

- **End users** (satisfaction, trustworthiness). For end users, explanations help them to better understand the features, qualities, and relevance of the items so as to make accurate next-step decisions (click, purchase, etc). In this case, a good explanation is supposed to be informative and useful to end users such that user satisfaction and trustworthiness are maximized.
- **System developers** (transparency, consistency). For those who develop recommender systems, explanations are mainly used to justify why the predictions are derived from the "black-box" models and hence to help developers to debug and scrutinize whether the model works as expected. In this case, the explanation is more measured towards transparency (whether the explanation can reveal model mechanism), and consistency (whether the explanation generation process is consistent with the actual decision-making steps).
- **Content providers** (effectiveness, efficiency). Content providers such as sellers and advertisers care more about how to improve effectiveness and efficiency by showing additional explanations with recommended items to users. The effectiveness of explanation can be measured through common metrics such as click-through rate, conversion rate, etc., while the efficiency depicts whether displaying explanations can expedite the end-user decision-making process.
- **Regulators** (scrutability). Regulators can leverage explanations to examine whether the recommender systems properly use users' sensitive personal data. Therefore, the measuring mechanism is similar to those for system developers where the resulting explanations are supposed to be transparent.

The explanation evaluation approaches can be generally divided into the following three categories, each of which faces a trade-off between result reliability and evaluation cost.

- **Offline evaluation**. It refers to using offline datasets and quantitative metrics to evaluate the quality of explanations generated by different explainable recommendation approaches. Depending on whether or not a ground-truth explanation is available and the forms of output explanations such as natural language explanations and instance-based explanations, the metrics can vary under different settings [61]. When a ground-truth explanation is available, we can use standard metrics to evaluate the extent that model generated explanation matches with the ground-truth explanation, such as precision, recall, and coverage for feature- or instance-based explanations [57, 170, 341], BLEU and ROUGE for natural language explanations [172], as well as NDCG, MRR and Hit Ratio for explanation ranking methods [171]. When ground-truth explanation is not available, we can use Probability of Sufficiency (PS) and Probability of Necessity (PN) to evaluate the sufficiency and necessity of explanations, such as evaluating counterfactual explanations [265, 266]. There is almost no cost to set up such evaluation process since it is equivalent to measuring model prediction on offline datasets.

- **User study and online evaluation**. This set of approaches recruit volunteers or select (a subset of) real users from commercial systems and manually evaluate the explanation quality through a simulated or real-world environment. The evaluation process could be either active (actively ask users some survey questions to evaluate the explanations) or passive (passively track and record some user reactions when interacting with explanations). For example, Zhang et al. [341] conduct A/B testing in a simulated environment over online users to evaluate how explanations influence users' click behaviors, while Xian et al. [304] conduct large-scale A/B testing in Amazon real-world commercial system to examine the effectiveness of increasing conversion and revenue when providing explanations in e-commerce recommendations. Online evaluation can better reflect users' reactions to explanations in real environments, but the experiment cost is also much higher and sometimes the evaluation environment is not accessible to researchers.

From existing research works, offline evaluation and survey-based user studies are widely adopted since they are doable and easy to set up, while online evaluation is less frequent since it requires access to the real-world system. We summarize common metrics adopted in these evaluation approaches as follows, which mainly depend on explanation forms. For natural language explanations, common metrics are BLEU and ROUGE [63, 120, 170, 172, 260, 314]. Besides, some outstanding work have extended and proposed several novel metrics in explanation evaluation. For example, Li et al. [170, 172] not only evaluate the generated sentence explanations but also evaluate how well the sentences really explain the recommendations based on newly designed metrics such as Unique Sentence Ratio (USR), Feature Matching Ratio (FMR), Feature Coverage Ratio (FCR) and Feature Diversity (DIV), which evaluates the uniqueness and personalization of the sentence explanations. For matching or ranking-based explanation (KG path, features, etc.), the common evaluation metrics are NDCG, Precision, recall and coverage [171, 235, 262, 263, 316]. For the evaluation of counterfactual explanation, common metrics are Average Treatment Effect (ATE), replacement, Probablity of Sufficiency (PS) and Probability of Necessity (PN) [56, 147, 265, 266, 275]. Some other evaluation metrics include Perplexity, Mean Explanation Precision (MEP), Mean Explanation Recall (MER) as well as influence [1, 187, 192, 224].

3.4 Open Problems and Relationship with Other Trustworthy Perspectives

- **Causal explanations**. Recently, there are several works [111, 147, 266, 275] which explore the counterfactual explainable recommender system from the causal perspective. However, finer-grained quantification of causal explanations is desired and yet to be explored, such as quantifying the causal treatment effect of the explanations and extracting the causally strongest explanations accordingly. On the other hand, there also exist works [177, 309, 311] which investigate the standard collaborating filtering problem from the causal perspective. However, less exploration has been conducted on utilizing causal learning frameworks such as confounding balance and do-calculus to better explain recommender systems from a causal perspective.

- **Controllable explanations**. Another important problem is how to make the explanation process more controllable, especially for generative explanation methods. Take textual sentence explanation as an example, existing methods derive user-readable textual explanations either by filling predefined templates [341] or directly generating the explanation sentence [169, 170, 172]. The previous approach is easy to control by filling the desired item features into the explanation template, but the template is fixed and not flexible enough for various different recommendation scenarios. The latter approach is very flexible and can automatically generate explanations based on the context, but it is usually not controllable in the sense that meaningless or even inappropriate text may be produced. One important direction is to explore how to control the content, quality, and flexibility of human-understandable explanations in the combined fields of controllable text generation [331] and explainable recommender systems.
- **Unbiased explanations**. The ultimate goal of explanation is to help users understand the model decisions and thus to gain trust. However, the explanation generation process may also be vulnerable to bias. For example, the explanation system may tend to generate some type of explanation for one group of people while generating another type of explanation for another group of people so as to persuade or even cheat humans to accept the model decisions, regardless of the real underlying mechanism that leads to the model decisions. If this is the case, explainable AI may completely violate the original goal—instead of gaining trust, it may actually lead to total distrust if users realize that the explanations are biased. To solve the problem, extensive efforts are needed to guarantee that explanations are faithful and unbiased. To date, seldom progress has been testified in the field of unbiased explanation, either identifying the underlying reason for model disparity or in recommendation [104] or exploring the trade-off between explainability and fairness/unbiasedness in trustworthy and responsible recommender system.

4 FAIRNESS

Recommender Systems have been considered as "benevolent" systems for a long time, which assist users (e.g., by helping them find relevant items) and create value for businesses (e.g., higher sales or increased customer retention) [141]. However, in the most recent years, considerable concerns from both academia and industry have been raised regarding the issue of fairness in recommendation [178]. Several studies argue that RS may be vulnerable to unfairness in several aspects, which may result in detrimental consequences for underrepresented or disadvantaged groups [109, 176, 180, 254]. For example, in e-commerce systems, RS may promote items that mainly maximize the profit of certain producers [103], or in online job marketplaces, RS may lead to racial or gender discrimination by disproportionately recommending low-payment jobs to certain user groups [109]. Therefore, to improve the satisfaction of different stakeholders in RS [2], it is important to study fairness in recommendation and build trustworthy and responsible systems.

4.1 The Source of Unfairness in Recommendation

Bias and discrimination are two common concepts related to the unfairness issue in machine learning and unfairness stems from both of these factors [178, 202]. In recommender systems, there are two main types of biases: bias in data, and bias in algorithm [178]. The biases in data may come from the processes of data generation, data collection, or data storage. For example, data bias may be caused by the specifics of the data collection process such as when a biased sampling strategy is applied. When training on the biased data, the recommendation models are highly likely to learn those over-represented groups, promote them in the ranked results, and potentially result in systematic discrimination and reduced visibility for disadvantaged user/item groups such as under-representing the minorities, certain racial groups, or gender stereotypes [176, 179]. Another source of unfairness may lie in the recommendation model itself. For example, recommendation

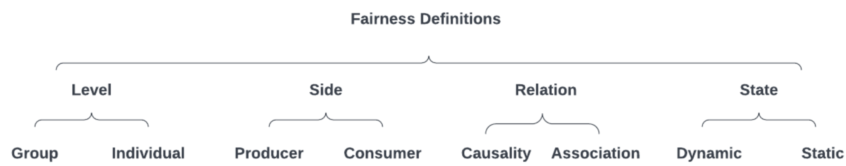

Fig. 3. Different dimensions to define fairness in recommendation.

models may further reinforce existing biases or existing skewed distributions in the underlying data. A well-known recommendation issue caused by such bias is popularity bias, where the popular items with more user interactions will be recommended more frequently and get more exposure opportunity than those less popular but equally or even more relevant ones [103]. Apart from bias, discrimination is caused by intentional or unintentional human prejudices and stereotyping with regard to sensitive attributes (e.g., race, gender, and religion, etc.) [202]. It is worth noting that bias and discrimination are not the only reasons of unfairness. For example, researches have shown that some fairness definitions cannot be satisfied simultaneously [68, 152, 226], as a result, the violation of one fairness definition may be caused by ensuring another fairness definition [178].

4.2 The Definitions of Fairness in Recommendation

Fairness is a comprehensive concept that can be defined on various different perspectives. An abundance of definitions has been explored in many research areas and tasks. Considering fairness in recommendation scenario makes the landscape even more convoluted. To clearly introduce the progress of fairness in recommender systems, we start from introducing its definitions. As fairness can be categorized into different classes from different perspectives, in this paper, we distinguish between individual and group fairness, consumer and producer fairness, associative and causal fairness, as well as static and dynamic fairness (Figure 3).

Fairness has been broadly defined as "the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making" [165, 202]. Therefore, in the first place, fairness is defined on group-level and individual-level.

- **Group fairness** is the idea that the average treatments should be the same across groups defined by certain attributes [279]. Such attributes are called protected (or sensitive) attributes, which often include gender, religion, age, sexual orientation and race, among others. Based on the above definitions, many variants are derived, such as Equal Opportunity, which requires that the true positive rate is the same across different groups [122], Equalized Odds, which requires that different groups should have the same true positive rate and false positive rate [122, 325], as well as Demographic Parity, which requires that each group should have the same likelihood to be classified as positive [42]. In recommendation scenarios, Li et al. [176] consider user-oriented group fairness. Specifically, they divide users into active and inactive groups based on the number of user interactions in the training data, and require that different user groups should receive similar recommendation quality such as F1 and NDCG.
- **Individual fairness** is the idea that individuals similar in their features should receive similar model predictions, namely, similar individuals should be treated similarly [85]. For example, Lin et al. [182] consider each individual user's utility of a recommended item as the relevance of the item to the user, and then considered fairness as the imbalance between users' utilities. Li et al. [179] study counterfactual fairness in recommendation, which is a type of individual fairness and requires that the recommendation results for each user should be same or similar in both the factual and the counterfactual world. The counterfactual world is defined as the one where user's sensitive features were changed while all the other features that are not causally dependent on the sensitive features remain the same.

Considering that the fairness demands in recommender systems may come from different stakeholders, fairness in recommendation is also divided into user (consumer)-side fairness and item (producer)-side fairness. Besides, there are also cases in which a system may require fairness for both consumers and producers, when, for instance, both users and items belong to protected groups.

- **User (consumer)-side fairness** studies the disparate impact of recommendation on protected classes of consumers [328]. The protected classes can be objective features such as race and gender, while it can also be subjectively assigned features. For example, Yao and Huang [319] investigate gender-based inequalities in collaborative filtering recommender systems, while Li et al. [176] group users based on their interaction frequency with the recommender and found that the active users accounting for a small proportion of the users (5%) enjoy much higher recommendation quality than others (95%). Even though most of the existing works focus on group-level user-side fairness, it can also be defined from individual-level. For example, Li et al. [179] leverage counterfactual fairness which requires that the recommendation results for each user are unchanged in the counterfactual world where the user's sensitive features are flipped, and each individual user can specify their sensitive features by themselves.
- **Item (producer)-side fairness** considers fairness for the items and the item producers in the recommender system, which ensures market fairness and avoids monopoly domination or Matthew's Effect [328]. For example, Ge et al. [103] and Abdollahpouri et al. [3, 4] focus on item popularity bias since popular items (i.e. those frequently rated, clicked or purchased items) get disproportionally more exposure while less popular ones are under-recommended. Moreover, some researchers explore producer fairness based on the sensitive attribute of the item producers such as their gender [33, 92, 110, 113, 151, 246], which is similar to the exploration of consumer fairness based on sensitive features.

The research community has studied fairness in machine learning by developing association-based (or correlation-based) fairness notions for a long time and most of the existing works about fairness in recommendations consider the association-based fairness notions. However, recently, some pioneering works have found that fairness cannot be well assessed only based on association notions [150, 160, 179, 332, 333], since such fairness definitions cannot reason about the underlying causal mechanism that leads to unfairness. As a result, we also introduce associative and causal fairness definitions.

- **Associative fairness** is also known as correlation-based/statistical fairness, which measures the statistical discrepancy between individuals or sub-populations, such as Equal Opportunity [122], Equalized Odds [23] and Demographic Parity [85, 326]. One weakness of associative fairness is that they cannot detect discrimination in presence of statistical anomalies such as Simpson's paradox, namely, the statistical conclusions drawn from the sub-populations could differ from that from the whole population [150, 160, 332].
- **Causal fairness** is not only based on data but also considers additional prior knowledge about the structure of the world in the form of a causal model. Besides, it is important to explore the causal relationship between the sensitive attributes and model output rather than just the associative relationship. In the recommender system domain, Li et al. [179] first explore counterfactual fairness in recommendation. The authors defined counterfactually fair recommendation as the recommendation results that are the same in the factual and counterfactual worlds for each user, where counterfactual world is the one where user's sensitive features are changed. For example, the gender of a user is changed from male to female, while all other insensitive features that are not causally-dependent on sensitive features remain the same. Such fairness guarantees that if

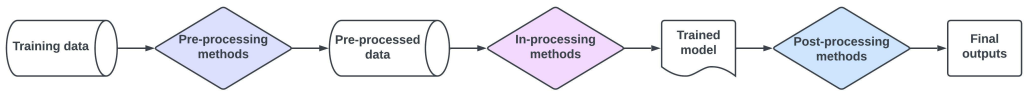

Fig. 4. Three types of fair recommendation methods in different parts of the recommendation pipeline.

the user does not want to be discriminated over certain sensitive features such as gender, then the user's recommendation results will be irrelevant to the sensitive feature.

Fairness in machine learning has predominantly been studied in static classification settings without concerning how decisions change the data over time. However, the fairness requirements in recommendation need to consider the dynamic nature of the systems since many features are changing over time such as user preference and item popularity.

- **Static fairness** provides a one-time fairness solution based on fairness-constrained optimization, which focuses on the fairness implications of decisions made in a static or one-shot context. Most existing fairness recommendation research put themselves in the static setting [178].
- **Dynamic fairness** considers the dynamic factors in the environment and learns a fairness strategy that accommodates such dynamics. For example, Ge et al. [103] study the dynamic fairness of item exposure in recommender systems. The items are separated into popular and long-tail groups based on the number of exposure in training data. The intuition of the work is that the item popularity may change during the recommendation process based on the recommendation strategy and user feedback, causing the underlying group labels to change over time, i.e., an item that was once unpopular may now become popular, and vice versa. To solve the problem, the authors formulate the problem as a Constrained Markov Decision Process (CMDP) with fairness constraint of item exposure change over time, and use Constraint Policy Optimization (CPO) to solve the formulated problem.

4.3 Methods for Fairness in Recommendation

Existing research on fairness in recommendation mostly focuses on three areas: 1) fairness quantification, which develops quantitative metrics to measure algorithm fairness under various fairness definitions, 2) fair recommendation modeling, which develops algorithms or models to improve fairness of the outputs, and 3) fairness diagnostics, which develops explainable fairness methods to identify the reason of model unfairness so as to explain why a model is fair or unfair.

Fairness quantification aims to develop and investigate quantitative metrics that measure algorithmic disparities in ranking or recommendation [96, 100, 176]. The works on fairness quantification have defined various types of unfairness in recommendations, such as unfairness in sensitive features like gender and age [55, 179, 319], unfairness in popular and unpopular items [3, 4, 103], unfairness in recommendation quality for users [96, 176]. For example, [96, 176] propose and studied the recommendation quality unfairness between active users and inactive users.

Fair recommendation models focus on providing fair recommendation results based on certain fairness definitions, which can be roughly divided into three categories: pre-processing methods, in-processing methods and post-processing methods [178, 180]. Figure 4 illustrates the differences between them.

- **Pre-processing methods** usually aim to minimize the bias in the source data before training the model. Therefore, they can be adopted when we have access to the data and usually they do not explicitly include fairness metrics that are defined over model outputs. Representative pre-processing methods include fairness-aware sampling techniques in the data collection process to cover items of all groups, balancing techniques to increase the coverage of minority groups, and repairing techniques to ensure label correctness [101]. For example, Lahoti et al. [161] introduce a method for mapping user records into low-rank representations that reconcile individual fairness through pre-processing. The method operates on an individual fairness objective to learn fair representations of training data points. The proposed method aims to transform an input feature vector into a fairer representation such that the individuals who are indistinguishable on their non-sensitive features in data should also be indistinguishable in their vector representations under a given distance function.

- **In-processing methods** aim to eliminate the bias during the model training process by modifying existing models or introducing new models. A general approach is to encode the fairness requirement as part of the objective function, typically as a regularizer [3, 25, 103, 176], which measures the degree of unfairness that the model must minimize in addition to the minimization of the original loss function. This approach also seeks to find a balanced trade-off between recommendation accuracy and fairness [103, 105]. In Ge et al. [105], the authors study the fairness-utility trade-off in recommendation scenario and proposed a fairness-aware recommendation framework based on Multi-Objective Reinforcement Learning (MORL), which is able to learn a single parametric representation for optimal recommendation policies over the space of all possible preferences (i.e., the Pareto frontier) between fairness and utility.
- **Post-processing methods** aim to modify the presentation of the already produced outputs to improve fairness through techniques such as re-ranking by linear programming [176, 254, 317] or multi-armed bandit [46]. For example, Zehlike et al. [327] propose a fair ranking algorithm—FA*IR to ensure that the number of protected candidates does not fall far below a required minimum percentage p at any point in the ranking. The method formulates this fairness as a statistical significance test of whether a ranking was likely to be produced by a Bernoulli process. Li et al. [176] propose a fairness re-ranking method which re-ranks the recommendation list of each user to guarantee fair recommendation quality for advantaged and disadvantaged users.

Fairness diagnostics focuses on answering a more fundamental question: what are the reasons that cause model unfairness? In general AI, there have been several pioneering works trying to derive explanations for model fairness [21, 221]. For example, Begley et al. [21] leverage Shapley value [247] to attribute the feature contributions to the model disparity so as to generate explanations. The proposed method estimates the sum of individual contributions from input features and thus understands which feature contributes the most to the model disparity [21]. Though this type of methods can provide explanations to the model disparity, they are not suitable for recommender systems due to the large item/user feature space in recommendation. To solve the problem, Ge et al. [104] design a learning-based counterfactual reasoning method to discover critical features that significantly influence the fairness-utility trade-off and use them as fairness explanations for black-box feature-aware recommendation systems.

4.4 Open Problems and Relationship with Other Trustworthy Perspectives

- **Fairness under other trustworthy perspectives**. Fairness is an important but not the only perspective for trustworthy recommendation and it is important to explore fairness under other perspectives. For example, explainable fairness: how can the algorithm explain to users or system designers why the model output is fair or unfair [104]; federated fairness: how can we simultaneously guarantee fairness and privacy in federated environment where global fairness constraints need to be considered in federated learning [189]; robust fairness: how can fairness-aware algorithms be resistant to malicious attacks that aim to make the system unfair. These research directions will enhance the trustworthiness of recommender systems from multiple dimensions altogether to make the system more reliable.

- **Long-term impact of fairness constraints**. Although some existing research shows that certain fairness and utility metrics may have to trade-off with each other in the short-term [105], we argue that understanding the benefits of fairness in recommendation should be considered in the dynamic and long-term context. This is because in the long-term, if users and producers feel they are being treated fairly by the system, their retention, interest, trust and engagement in the system will increase, which will in turn help to create and nourish a sustainable economic eco-system in the platform.
- **Controllable Fairness**. It has been theoretically proven that some fairness notions are inherently conflicting with each other and cannot be achieved at the same time [152]. Therefore, to construct a controllable fairness system, where the users and producers can select the kinds of fairness they care about most, becomes a feasible solution. There have already been several pioneering works focusing on a similar idea, such as [179, 298], however, they mainly focus on user-side controllability and ignore the producer-side.

5 PRIVACY

With the growing concerns about the machine learning methods that gather and analyze personal data, the ethical demand for data privacy has been formally recognized in terms of mandatory regulations and laws [222, 281]. As a consequence, the research of privacy-preserving machine learning has witnessed substantial development in recent years [185]. It is believed that a more trustworthy web service would provide privacy-protected solutions that avoid unwanted exposure of information for any participants of the system.

In both the recommender system and the general machine learning field, there exist several definitions of privacy [5, 148, 255, 257], and in most cases, they share the same ingredients:

- **Private information**: the critical or valuable information that needs restriction of access. For example, user identity and sensitive user attributes such as gender, age, and address.
- **Ownership**: only the authorized entities can access and control the corresponding private information, where the entity may refer to a user or even the platform itself.
- **Threat**: malicious entities (inside or outside the system) aiming to get access to or manipulate private information. Note that such entities may utilize auxiliary *public information* to engage its infiltration or attacks.
- **Goal of privacy protection**: to maintain the *ownership* of the *private information* and find countermeasures for the *threats*.

In this section, we adopt these terminologies and discuss the privacy problems in the field of recommender systems (RS). We first explain the privacy demands of different types of ownership in RS in Section 5.1, and then list the major privacy threats and challenges in Section 5.2. Section 5.3 illustrates several major privacy-protection techniques. And finally, we enumerate several open questions in Section 5.4. For existing surveys, we consider [95] as the backbone of our taxonomies, then include and extend the ideas from several recent surveys [5, 136, 282].

5.1 Ownership of Private Information

While the term "privacy" is used in a variety of scenarios on the Web, privacy problems in RS are typically related to two types of entities: users/customers and the recommendation platform itself. Each type of entity corresponds to a specific type of ownership demand and thus faces different privacy risks.

User Privacy: To provide accurate personalized predictions, recommender systems heavily depend on rich user digital traces. This may involve the collection of sensitive or critical user information and potentially threaten users' privacy. Recent debates and regulations [281] on user's data

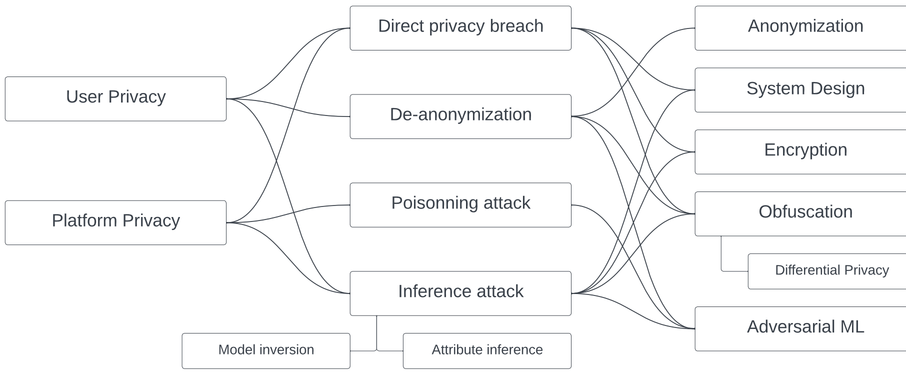

Fig. 5. Ownership types (left), privacy threats (middle), and protection techniques (right).

privacy brought up concerns about platforms gathering sensitive user information such as gender, race, age, location, sexual orientation, and contact information, and potentially using it in unlawful or untrustworthy ways, such as targeted advertising and scamming. In RS, this is a natural conflict since most recommendation solutions are personalized and require such information to construct accurate user profiles. Though some research found that users may become less concerned about privacy leakage to trusted services [72], the trade-off between recommendation personalization and user privacy has been one of the critical design challenges for modern recommender systems [15, 175].

Platform Privacy: Even if the recommendation platform rightfully collects and utilizes the data from users, there are still privacy risks if outside attackers infiltrate the system or reveal critical system information such as the log data and model parameters. On one hand, when the RS does not provide sufficient system security and data anonymization, adversaries may recover the user's data and system information through hacking or inference [41]. On the other hand, one can pretend to be a user or a trusted third-party, and interact with the RS in order to influence the decision system. For example, a "fake user" can inject disguised data that is specifically designed to trick the RS, so that the resulting recommendation model learns to favor certain items or business owners and suppress the others [90].

In this survey, we address the difference between these two types of ownership from the perspective of trust management: while the user privacy problem mainly concerns the users' trust towards the RS platform, the platform privacy discusses the system's trust towards its users and external entities. In general, the RS should show both its legitimacy in collecting users' data and its ability to protect its data and model against privacy attacks. For the following sections, we first discuss existing research works in terms of the privacy threats and then list several open questions in the field according to this taxonomy.

5.2 Privacy Threats

In this section, we summarize existing research about privacy threats in recommender systems and categorize them with respect to each type of ownership category. The taxonomies are provided in Figure 5. We denote "privacy breach" as the efforts that aim to directly access private information through hacking or monitoring. This type of attack is mainly based on prior knowledge of the software implementation details of the communication interface (i.e. the APIs provided by the RS), which is out of the scope of this survey. We illustrate the details of each of the remaining topics as follows.

Deanonymization. Recommender system may disclose its data to third parties or to the public, e.g. e-commerce services open their consumer log data to researchers. Since the data in recommender system typically includes comprehensive user information, the system has to protect users' Personally Identifiable Information (PII) in the data through anonymization, protecting user information [158, 212]. Such information includes but is not limited to the user ID, user interaction history, and user profile features such as gender, age, and address. In this context, the PII is user's private information and the full ownership should belong to the user. However, studies have shown that one can still de-anonymize (or re-identify) user identities by 1) cross-linking them with external information [98, 168, 210], or 2) inference from the partial observation of the same sample as we will explain later. For simplicity, we denote de-anonymization as the first scenario which requires attribute linking with external data, while the latter is denoted as inference attack.

Inference attack. Users or the recommender system platform may deny the upload or sharing of critical information to physically maintain privacy. However, studies have shown that it is still possible to accurately infer the protected attributes merely by using public information. In terms of the user privacy, works have shown that the user's private information can be inferred by their behavior history [26, 41, 114], rating data [292], user interests [47], social relations [81, 124], or even the writing style of the textual feature [234]. This is also referred to as the attribute inference attack, and the attacker could be either the RS or external entities that have obtained public information about the user and the RS. In terms of the platform privacy, the existence of the recommendation also opens the door to the model inversion attack [93], where attackers can observe the behavior of the model and then reconstruct the private information. For example, by sending queries and receiving feedback from the RS, one can reconstruct the missing private attributes of a sample [130, 201], infer if a certain record (or user) is present in the training data of the model [251, 334] (a.k.a. membership inference attack), and even recover the model parameters [274]. In general, the inference attack does not always require auxiliary data with correspondence of attributes like that in de-anonymization but always relies on the data correlation or the predictive model to engage the inference.

Poisoning attack. This type of threat aims to indirectly control the recommendation model (regarded as private information) by modifying the training data through the provided trusted channel. Compared with the previously mentioned attack methods that "read" the private information, the poisoning attack distinguishes itself for it aims to "writes" the private information. The most typical example in RS is the fake user attack [90, 218] (also known as profile injection attack [37] or shilling attack [44, 162, 253, 342]), and it creates fake users to interact with the system for certain entities' advantage. Another type of poisoning attack is the co-visitation injection attack [315], which aims to inject target items into the user viewing history by modifying the recommendation list before showing it to users. In general, the poisoning attack can be regarded as an optimization problem and it is also related to the problem of RS robustness as we will show in Section 6. It usually requires the knowledge of the training data so as to figure out a way to inject the data that can serve the "attack purpose" without significant change in the distribution, making the attack hard to detect. Such "attack purpose" includes 1) reducing the overall RS performance, 2) changing the overall rating of some target item, and 3) raising or suppressing the chance of recommendation of certain items. Additionally, studies have shown that poisoning attacks can be detected to some degree [28, 38, 39, 342, 344], though there are always chances that the detection method produces false positives that significantly influence the service quality and false negatives that overlook a successful attack. One specific type of attack that is closely related to poisoning attack is the adversarial attack [76, 78, 167] which constructs a set of instances that aim to make the recommendation model inaccurate. Differently, there are poisoning attacks that are non-adversarial (i.e. raising a certain item's rating), and there are certain adversarial attacks that do not poison the data [191], which is usually a robustness problem rather than a privacy problem.

5.3 Privacy Protection Techniques

Anonymization. The confrontation between anonymization and deanonymization has been around for years, and the goal is to hide certain user details (i.e., PII) when publishing the dataset to third parties. There are several studies proposed to solve this privacy issue based on the estimated risk of identifying a user in a system, such metrics include k-anonymity [58, 261], l-diversity [195], and t-closeness [174]. Another effective solution is to replace the detail with high-level group information (i.e., through clustering) [269]. As described in the previous section, simply replacing attributes with pseudonyms does not guarantee privacy since one can still de-anonymize the private information even with a little bit of auxiliary background knowledge [98, 168, 210]. Thus, perturbation-based approaches such as differential privacy are introduced to alleviate this problem as we will illustrate later in this section.

System Design. This type of solution is one of the earliest works that discuss the system-level privacy issue. Some typical insights include 1) letting users know about the privacy risks and acquire consent, 2) certification and verification between multiple parties [70], 3) dynamic authentication and temporal limits for access, and 4) risk reduction by distributed data storage and computation. The last solution with distributed data can be further categorized into the service-side distribution and the user-side distribution. The service-side distribution involves multiple functional participants that jointly provide the recommendation service [8], and the user-side distribution (e.g. federated learning or block-chain) allows users to keep their own data on personal devices without uploading to the central server [43, 48, 145, 209, 219, 295].

Encryption. The RS communicates with user devices and third-party services, as a result, it introduces the risk of data leakage (i.e. through external monitoring or hacking) whenever there is message transfer. Thus, encryption techniques are used to encode the transferred information so that no one except for the owner can easily decipher the message [43]. One widely adopted method is the homomorphic encryption [41, 48, 87, 256, 329], which allows third party computation without decryption or computational error. Another method that provides secure computation is the garbled circuits with public-key encryption [211], where the collaborative filtering model is encrypted as a circuit, and multiple parties jointly optimize the model without revealing their input data. In general, the encryption techniques are often used as fundamental building blocks for multi-party and communication-rich frameworks such as federated learning and secure multi-party computation [32]. Though these approaches provide accurate computation and guaranteed privacy, it also induces non-trivial extra computational cost [16].

Obfuscation and Differential Privacy. When the leakage of the data (either through intentional publication or unintentional leakage) is inevitable, an alternative solution is adding noise to the data so that it can disguise the actual value. This type of solution is effective in RS since the recommendation model can usually be expressed as an "aggregation" of individual information. Intuitively, the introduced noise should significantly change the value of each record (or user) such that the actual value is hard to infer, and the noise should also be sufficiently small such that the aggregated value is still statistically accurate. Early works studied simple obfuscation and perturbation methods [227, 292], and later approaches formulate this as differential privacy [24, 64, 200, 348]. Compared to encryption methods, these noise-injection solutions are more efficient in practice but introduce additional computational error and utility loss.

Adversarial Machine Learning. One critical problem of obfuscation-based methods is the loss of recommendation utility, as a result, some recent works propose to learn special noises that achieve differential privacy without compromising the utility by formulating the noise finding task as a machine learning problem [50, 144]. Note that this type of solution also belongs to the family of differential privacy methods and it is specifically designed for defending against user attribute inference attacks. A different type of adversarial machine learning solution is training an adversarial model that mimics the attackers' behavior for a wider range of attacks [13, 22, 126]. This adversarial model is later utilized to train an adversarial-aware recommendation model that is robust against such attacks without significant loss of recommendation utility. Both types of adversarial learning are closely related to the system robustness, so it is particularly suitable for fighting against poisoning attacks and inference attacks.

5.4 Open Problems and Relationship with other Trustworthy Perspectives

Privacy in decentralized systems. Recent developments on Internet of Things (IoT) systems and blockchain have brought new ideas for further distributing the control and computation towards edge devices. Providing recommendation services in such systems can systematically reduce privacy risks because of the distributed data storage and model computation [80, 94, 115, 184]. However, it also indicates that the central service will have weak or even no control over the RS. Different from the centralized RS which only involves communications between a user's edge device and the central server, decentralized RS involves direct communication between edge devices, which introduces new privacy challenges yet to be explored.

Explainable AI and Privacy. Some existing privacy protection techniques are built based on deep learning models that are difficult to interpret. However, it is sometimes highly demanded to explain how the system guarantees privacy, why it is effective against certain privacy risks, and what are possible side effects. This topic is one of the new research areas in explainable AI and it is closely related to the research of explainable security that recently gained research attention [205, 242]. Besides, privacy attacks may target the system's explanation model, since the explanation model may reveal critical information about the underlying predictive model [159]. As a result, it is important to explore privacy-preserving explanation models that explain the model decisions and meanwhile protect system security. Both topics are not yet discussed in RS research.

Conflict with Fairness. To protect user privacy, the system especially needs to protect users' sensitive attributes such as age, gender, race, and location and avoid revealing such information to any party. However, such sensitive attributes are important and usually highly needed by algorithm developers to develop fairness-aware models so as to improve fairness treatment of users. Such algorithm developers may include the developers within the company, contracted third-party developers, or even non-contracted independent developers. As a result, there may exist an intrinsic conflict between promoting privacy and promoting fairness, since promoting fairness requires access to sensitive features while promoting privacy refrains from sharing them, which has been pointed out as a critical challenge for privacy-protected machine learning [347]. In some cases, the attributes that require fairness control (e.g. gender) are exactly the sensitive information that needs privacy protection [252]. Thus, it is more important in such cases to find a solution that can achieve fairness without affecting privacy. Additionally, studies have also shown that certain privacy protection methods may naturally improve fairness since they essentially disguised the private information during prediction [239]. However, it is still unknown whether there are certain connections between the privacy protection methods and the system's fairness and what are the cause and effects.

6 ROBUSTNESS

Though recommender system promotes the efficiency of information seeking and benefits both customers and producers, it may also expose the users of the system to threats in terms of robustness, which leaves space for third parties to manipulate recommendation outcomes to users through profile injection attacks (a.k.a shilling attacks). The motivation for such attacks is often malicious, e.g., personal gain of illegitimate profits, market penetration of certain items/brands, or even causing malfunction of the system. Since recommender systems have already been adopted in many high-stake decision-making scenarios, such vulnerability raises concerns regarding to how machine learning techniques can be safely adopted in recommender systems, and how could they be carefully designed so as to be robust and trustworthy against aggressive provocation from attackers [13].

6.1 Taxonomy of Attacks

Attacks against recommendation models can be categorized according to various dimensions [13, 78, 207, 253], such as the timing of the attack, the intent of the attack, the size of the attack, and the knowledge of the attack, among others.

- **Attack Timing**. Based on when the attacks occur in the learning pipeline, they can be categorized as poisoning attacks and evasive attacks. On one hand, poisoning attack happens before the model is trained and the attackers add attacking data points into the training data, causing the trained model to make erroneous predictions. As a result, poisoning attacks can directly influence the trained model. On the other hand, evasive attacks happen after the model is already trained and thus do not influence the trained model itself, instead, it aims to inject fake results into the model output while avoiding being detected by the model [13, 29].
- **Attack Intent**. Different attackers may have different intents. "Push attack" and "nuke attack" are two basic intents, where an attacker might insert fake profiles into RS to make an item/producer more likely to be recommended ("push") or less likely ("nuke") by the recommendation algorithm [13]. Apart from these, the attacker may even attack the system just to make it recommend irrelevant items to users so as to lower users' trust.
- **Attack Size**. The size of an attack can be measured in several ways. The most frequently used measure is the number or percentage of profiles injected into the system by the attacker [294]. Statistics indicate that in many commercial recommender systems, the total quantity of injected profiles is usually around 1–15% due to the level of effort and information needed to successfully execute the attack [253].
- **Attack Knowledge**. A high-knowledge attack is one that requires very thorough knowledge of the data distribution in a recommender system's database, and if the attackers reproduce the precise details of the data distribution within the profile database, then this attack is called a perfect-knowledge attack [293]. Furthermore, a low-knowledge attack is one that only requires system-independent knowledge such as that obtained by consulting public information sources [13, 207].

Based on existing literature, there are two main classes of attacks on RS: 1) hand-engineered shilling attacks, and 2) machine-learned adversarial attacks. The former relies on hand-engineered fake user profiles (typically a rating profile) injected into the system, while the latter is machine-learned attack optimized to find minimal perturbation of the user-item rating matrix or user/item content data to influence the recommendation performance. Moreover, machine-learned attacks can also be seen as a novel type of shilling attack that applies the adversarial learning paradigm for generating poisoning input data.

6.2 Hand-engineered Shilling Attacks

Attack. Shilling attacks against RS have established literature. Since the early 2000s, the literature was focusing on building hand-crafted fake profiles whose rating assignment follows different strategies, such as random [163], popular [213], love-hate [207], bandwagon [214], and average [163], among others. Specifically, given a user-item interaction matrix, the goal of a shilling attack is to insert fake profiles into the matrix in order to affect the predicted ratings and/or diminish the performance of the system to reveal the attackers' engineered and illegitimate targets, e.g., pushing some targeted items into the top-K recommendation list to improve their market penetration [118].

Defence. Generally, there are two commonly-used methods to mitigate the impact of shilling attacks. One method is to use a shilling attack detection algorithm and then remove the detected attack profiles from the rating matrices before training the recommendation algorithms. There have already been a large number of works developed to detect such shilling attacks [27, 118, 253, 318, 342], and such methods can be roughly classified into supervised classification methods and unsupervised clustering methods. The majority of work on *supervised classification methods* begins with feature engineering and then moves on to algorithm development. For example, Yang et al. [318] extract three well-designed features from user profiles to identify attack profiles, namely, the filler size with maximum, minimum, and average ratings, where filler size is the ratio between the number of items rated by the user and the number of entire items in the recommender system. Then, the features are analyzed with statistical tests and classified using a variant of AdaBoost method named the re-scale AdaBoost (RAdaBoost) based on extracted features to proceed with the attack profile detection. *Unsupervised clustering methods* usually aim to group individuals into groups and then eliminate suspicious ones. For example, Bhaumik et al. [27] use k-means clustering to detect relatively small clusters as attack groups of user profiles, while Zhang et al. [342] develop a propagation method based on matrix iteration to calculate the likelihood of each user to be an attacked user based on a small seed group of known good users. Another alternative defense option is to develop attack-resistant recommendation algorithms, which is trying to reduce the influence of shilling attacks. For example, in [330], Zhang et al. propose a robust collaborative filtering method by incorporating non-negative matrix factorization (NMF) with R1-norm, while in [320], the authors' design a robust matrix factorization model based on kernel mapping and kernel distance.

6.3 Machine-learned Adversarial Attacks

The research works focusing on machine-learned adversarial attacks against RS have recently received considerable attention from the research community [45, 190]. Based on their level of granularity, researchers mainly classify machine-learned adversarial attacks into three categories: 1) adversarial perturbation of model parameters, 2) adversarial perturbation of contents, and 3) machine-learned data poisoning attacks [13, 79]. We introduce each of these attacks as well as the corresponding methods to defence them.

- **Adversarial perturbation of model parameters**. *Attacks:* The primary goal of this type of attack is to add adversarial perturbation on the user and item representations in the latent space. For example, He et al. [126] study adversarial attack strategies against the recommender system's parameters with a specific focus on the robustness of BPR-MF [236] against adversarial perturbations on the user and item embeddings. Specifically, the authors first generate the perturbations based on the Fast Gradient Sign Method (FGSM) proposed by Goodfellow et al. [116], then add this perturbation to the model parameters, and finally generate the recommendation list with this perturbed model parameters. After exposing the parameters of BPR to both adversarial and random perturbations, they claim that the value of Normalized Discounted Cumulative Gain

(NDCG) is decreased by -21.2% and -1.6%, respectively, which shows a huge difference. Several other recent works have performed similar FGSM perturbation against different recommender systems, such as visual-based recommender [267], factorization machines [50], deep neural network models [323] and auto-encoder models [322]. *Defense:* To protect recommenders against FGSM attack, [126] make the model robust to parameter perturbations by modifying the classical loss function of a recommender model (i.e., BPR-MF) by adding an adversarial regularization term. The application of adversarial training has been shown to improve the model's robustness against FGSM attacks, and then this training process has been applied to a variety of recommendation models, such as BPR-MF [126] and VBPR [267].

- **Adversarial perturbation on content data**. *Attacks:* This type of attacks focuses on the perturbation of content data related to users and items, and it is mainly applied to content-based and hybrid recommender systems. For example, Gao et al. [99] propose a black-box attack strategy, called DeepWordBug, to effectively generate a spam movie review that fools a deep RNN-based classifier to misclassify it as a positive message. *Defense:* Anelli et al. define three main defence methods, which include: 1) robustification procedure, such as deep contractive network [117], data compression [86] and data randomization [306], 2) detection techniques of adversarial samples, which detect the adversarial samples by extracting the content features based on machine learning methods, such as principal component analysis (PCA) [67, 203, 204], and 3) increasing model robustness to adversarial samples, such as the adversarial training of recommendation algorithms by adversarially regularizing the recommendation model [51, 84, 126, 267, 323].
- **Machine-learned data poisoning attacks**. *Attacks:* Similar to previously mentioned poisoning attacks, this class of attacks targets the learning algorithms by manipulating the data used for training these models, which happens before the recommendation model is trained. The difference is that they are learned through specific optimization procedures maximizing the adversary's goal automatically. For instance, Fang et al. propose a relevant work [89] tackling the data poisoning optimization for top-N recommendation, where the attacker's goal is to promote a particular item to as many normal users as possible and maximize the hit ratio, which is defined as the fraction of normal users whose top-N recommendation lists include the target item. In particular, inspired by interpretable machine learning, the authors exploit the influence function approach, which claims that the top-N recommendations are mainly affected by a subset S of influential users, then they propose a gradient-based optimization algorithm based on S, named S-attack, to determine the fake users' rating scores, and finally add the fake users to the recommender system. *Defense:* The defense mechanisms against this kind of attacks can be also classified as detection methods and methods seeking to increase model robustness. For the former, previous studies extracted several features from rating scores to train a binary classier to distinguish between normal users and fake users [90], i.e., Rating Deviation from Mean Agreement (RDMA), Weighted Degree of Agreement (WDA), Weighted Deviation from Mean Agreement (WDMA), Mean-Variance (MeanVar), Filler Mean Target Difference (FMTD). For the latter, Hidano and Kiyomoto [129] propose a defensive strategy leveraging trim learning, which exploits the statistical difference between normal and fake users/items, to make matrix factorization resistant to data poisoning.

6.4 Open Problems and Relationship with other Trustworthy Perspectives

- **Understanding the reasons of (un)robustness**. Though researchers have developed various methods to detect and improve the robustness of recommendation models, it is still difficult to understand why a model is robust or unrobust, and what reasons lead to the robustness problems of a model. To solve the problem, it is important to explore explainable robustness methods, since understanding the underlying reasons for (un)robustness can help system developers to better maintain the model robustness and resist attacks.

- **Effects beyond accuracy metrics**. Most of the research on adversarial machine learning and recommender system tends to focus on accuracy metrics, however, the impact on other metrics beyond accuracy could also be the main objective of a new class of attack strategies aiming to compromising the beyond accuracy metrics of RS, such as the fairness, explainability, and privacy of the recommendation results. As a result, attack defense algorithms should not only care about defending the recommendation accuracy, but also defending a broader scope of metrics.
- **Attacks beyond user-item matrix**. The majority of the proposed attacking or defending strategies mainly deal with the collaborative data available in the user-item matrix, whereas modern recommendation models utilize a large amount of auxiliary information beyond that [249], such as images, videos, texts, social connections, etc. Therefore, how to develop adversarial attacks against these heterogeneous data remains an open problem.

7 CONTROLLABILITY

Controllability of AI is one of the most important problems facing humanity [313], which is essential when users interact with intelligent systems and has been studied in the human-computer interaction (HCI) community for over two decades [11, 276]. In recommender systems, which actively interact with humans [154, 199, 245, 250, 258], the importance of controllability can not be neglected. However, despite the recent successful improvement of the recommendation performance, the controllability issue in recommender systems has become a new major concern: most of the current RS are mostly uncontrollable by the system user and users can only passively receive the recommendation results. More specifically, when using a non-controllable recommender system, users can only passively choose to accept or not accept the recommendation results, but they can hardly control what the recommendation results they receive, and more importantly, what the recommender systems learn about them. In fact, controllability is an important aspect for building trustworthy recommender systems. Recent studies have shown that users may not be satisfied even with a high recommendation accuracy [128, 198], and increasing the users' controllability over recommender systems can increase the users' satisfaction and trust in the recommendation results [133, 142, 146, 153, 197, 305, 340].

Recommender systems provide personalized recommendations based on user preferences, which are learned from users' interaction history. No matter what techniques are used for recommendation, the user preference construction process is crucial for making personalized recommendations [14]. However, traditional non-controllable RS fails to solve two problems: 1) what if the constructed user preferences are not accurate, and 2) even with accurate user preferences, what if the users are not willing to receive recommendations based on all of their preferences and information. User controllable recommendation tackles the above problems by allowing users to manually express their preferences and intervene in the preference construction process via certain types of interactions. In this section, we introduce the existing research on user controllable recommendation, which can be roughly classified into two categories according to the two different types of interventions: 1) explicit controllability based on explicit intervention, and 2) implicit controllability based on implicit intervention. Then, we discuss the open problems to be studied in the future. We note that Jannach et al. [143] conducts a concrete survey on the topic of controllable recommendation. Compare to [143], in this section, we use a different taxonomy and focus more on the research conducted in recent years.

7.1 Explicit Controllability

Research on controllable recommender systems based on explicit controllability let the users explicitly edit or update the user preferences. These systems assume that the users are not only aware of their desires and favors, but able to accurately express them. In these designs, users are knowledgeable or at least partially knowledgeable about how their feedback affects the construction of user preference. The most common approach is to let users set their profiles or reweight the pre-defined aspects or features to directly intervene in the preference construction process. For instance, Hijikata et al. [132] propose a content-based filtering system for music recommendation using a decision tree, which enables users to edit the learned profiles on the tree. Similarly, in [133], users are allowed to control a music recommender system by selecting the categories of the preferred items or editing their profiles. Terveen et al. [270] propose a graphical interface model in a music recommender system, where users can stretch the bars in a preference histogram to increase or decrease their intent on different specifications. Knijnenburg et al. [154] propose an interaction strategy where users can explicitly assign weights to each of the item attributes for controllability. Later in [34], a more complicated hybrid music recommendation system is developed, where users can move sliders associated with the weights assigned on various items or features to observe the changes on the recommendation list. Other works combine scrutability with profile setting [164, 289]. In [289], a scrutable recommender system is used to recommend meals at a particular restaurant, where users are notified about how the preference contributes to the meal's personalization score. In [164] and [18], textual, visual or attribute-based explanations are provided for users in the recommender system. In these works, if users are not satisfied with the recommendation list, they are able to modify the user preferences accordingly.

7.2 Implicit Controllability

Controllable recommendation methods based on explicit controllability face several issues. For example, the users may not always be fully aware of their preferences. The users may also be concerned about the effect or consequences of their revealed preferences due to the intransparency of the recommendation models. Therefore, more recent works provide implicit controllability to the users. The crucial idea of implicit controllability is that the users do not directly manipulate the user profiles or favored features. Instead, they indirectly fine-tune their preferences when dynamically interacting with the recommender systems. For example, Harper et al. [123] allow users to interact with the recommendation results, in which users can re-rank the generated recommendation lists to fit their expectations. After re-ranking, the system will automatically update the users' preferences by re-weighting on two attributes: popularity and age of the items. Schaffer et al. [244] let users interact with the information used by the system to construct their preferences. In their developed interface, users are able to add and delete their historical interactions with items or modify the rating on the past items to observe the changes in recommendation results.

7.3 Open Problems and Relationship with other Trustworthy Perspectives

As research on user controllable recommendation continues to explore the feasible control options for the users, which improves the users' trust and satisfaction in the recommender systems, there are still open problems to the studied in the future. We list some of the important directions to be studied in the following.

- **Explainability of controllability**. While more and more interaction options are provided to the users to increase their controllability and satisfaction, controlling the AI systems without understanding the underlying mechanism may cause errors or even harm than benefits [19]. In most of the existing works on controllable recommender systems, no matter whether the controllability is provided explicitly or implicitly, users only acquire an interface to intervene in the recommender system without knowing the consequences: in explicit controllability, users can directly interact with their preference information, but they do not understand how the preference contributes to the personalized recommendation results; in implicit controllability, users change the recommendation results to the desired direction so as to implicitly update their preference. However, they are not knowledgeable of how their interactions will influence their preference. Some research such as [164, 289] explains to the users which parts of their preference contribute to the recommended items before letting users change their preference, which may be a good attempt toward explainable controllability but is still in a very early stage.

- **Personalized controllability**. Since users have different preferences, modern recommender systems provide users with personalized recommendations. However, this is not the case in the current research on controllability. Existing controllable recommender systems provide users with exactly the same control options with no personalization. In the future, it is important to consider that different users may have different preferred types and degrees of controllability and develop personalized controllability techniques for recommender systems.
- **Evaluation of controllability**. Last but not least, evaluation methods for controllability are an important direction to explore. In existing research, when a new controllability method is proposed, researchers usually use user studies as the most common method to evaluate the proposed methods. So far, we still do not have a generally accepted quantitative evaluation metric to evaluate the various controllability methods in a unified way. A faithful evaluation framework should be able to measure different aspects of controllability, such as how powerful the controllability is for users or what is the cost of this controllability. The evaluation method is an important open problem to be explored in the future for user controllable recommendation.

8 CONCLUSION

This survey summarizes the current developments and trends in trustworthy recommender system research, with the goal of facilitating and advancing future research and implementation of trustworthy recommender systems. From a technological point of view, this survey provides a road map for comprehensively developing trustworthy recommender systems. We begin this survey by defining trustworthiness for recommender systems and illustrating their characteristics by categorizing trustworthiness principles. Following that, we introduce and discuss the recent advances in trustworthy recommender systems in terms of explainability, fairness, privacy, controllability, and robustness. We describe the fundamental ideas for each component, provide a detailed overview of existing methods for each of them, and then suggest the prospective future research directions for these elements, especially from cross-aspect perspectives. Overall, the research field of trustworthy recommender systems is important and booming with a diverse set of approaches and applications, meanwhile, making recommender systems responsible and worthy of our trust is one of the biggest challenges that our research community needs to combat. We hope the survey equips researchers interested in this area with sufficient background and knowledge to meet this challenge.

REFERENCES

[1] Behnoush Abdollahi and Olfa Nasraoui. 2017. Using explainability for constrained matrix factorization. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 79–83.
[2] Himan Abdollahpouri and Robin Burke. 2019. Multi-stakeholder recommendation and its connection to multi-sided fairness. *arXiv preprint arXiv:1907.13158* (2019).
[3] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the eleventh ACM conference on recommender systems*. 42–46.
[4] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286* (2019).
[5] Erfan Aghasian, Saurabh Garg, and James Montgomery. 2018. User's Privacy in Recommendation Systems Applying Online Social Network Data, A Survey and Taxonomy. *arXiv preprint arXiv:1806.07629* (2018).
[6] HLEG AI. 2019. High-level expert group on artificial intelligence. , 6 pages.
[7] Qingyao Ai, Vahid Azizi, Xu Chen, and Yongfeng Zhang. 2018. Learning heterogeneous knowledge base embeddings for explainable recommendation. *Algorithms* 11, 9 (2018), 137.

[8] Esma Aïmeur, Gilles Brassard, José M Fernandez, and Flavien Serge Mani Onana. 2008. Alambic: a privacy-preserving recommender system for electronic commerce. *International Journal of Information Security* 7, 5 (2008), 307–334.

[9] Ahmad Al-Doulat. 2021. *FIRST: Finding Interesting StoRies about STudents-An Interactive Narrative Approach to Explainable Learning Analytics.* Ph.D. Dissertation. The University of North Carolina at Charlotte.

[10] Kars Alfrink, Ianus Keller, Neelke Doorn, and Gerd Kortuem. 2022. Tensions in transparent urban AI: designing a smart electric vehicle charge point. *AI & SOCIETY* (2022), 1–17.

[11] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems.* 1–13.

[12] Marina Andric, Iustina Ivanova, and Francesco Ricci. 2021. Climbing Route Difficulty Grade Prediction and Explanation. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.* 285–292.

[13] Vito Walter Anelli, Yashar Deldjoo, Tommaso DiNoia, and Felice Antonio Merra. 2022. Adversarial recommender systems: Attack, defense, and advances. In *Recommender systems handbook.* Springer, 335–379.

[14] Müslüm Atas, Alexander Felfernig, Seda Polat-Erdeniz, Andrei Popescu, Thi Ngoc Trang Tran, and Mathias Uta. 2021. Towards psychology-aware preference construction in recommender systems: Overview and research issues. *Journal of Intelligent Information Systems* (2021), 1–23.

[15] Naveen Farag Awad and Mayuram S Krishnan. 2006. The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS quarterly* (2006), 13–28.

[16] Shahriar Badsha, Xun Yi, and Ibrahim Khalil. 2016. A practical privacy-preserving recommender system. *Data Science and Engineering* 1, 3 (2016), 161–177.

[17] Marko Balabanović and Yoav Shoham. 1997. Fab: content-based, collaborative recommendation. *Commun. ACM* 40, 3 (1997), 66–72.

[18] Krisztian Balog, Filip Radlinski, and Shushan Arakelyan. 2019. Transparent, scrutable and explainable user models for personalized recommendation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval.* 265–274.

[19] Gabriel Diniz Junqueira Barbosa and Simone Diniz Junqueira Barbosa. 2020. You should not control what you do not understand: the risks of controllability in AI. *Human Computer Interaction and Emerging Technologies: Adjunct Proceedings from* (2020), 231.

[20] Jordan Barria Pineda and Peter Brusilovsky. 2019. Making educational recommendations transparent through a fine-grained open learner model. In *Proceedings of Workshop on Intelligent User Interfaces for Algorithmic Transparency in Emerging Technologies at the 24th ACM Conference on Intelligent User Interfaces, IUI 2019, Los Angeles, USA, March 20, 2019*, Vol. 2327.

[21] Tom Begley, Tobias Schwedes, Christopher Frye, and Ilya Feige. 2020. Explainability for fair machine learning. *arXiv preprint arXiv:2010.07389* (2020).

[22] Ghazaleh Beigi, Ahmadreza Mosallanezhad, Ruocheng Guo, Hamidreza Alvari, Alexander Nou, and Huan Liu. 2020. Privacy-aware recommendation with private-attribute protection using adversarial learning. In *Proceedings of the 13th International Conference on Web Search and Data Mining.* 34–42.

[23] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44.

[24] Arnaud Berlioz, Arik Friedman, Mohamed Ali Kaafar, Roksana Boreli, and Shlomo Berkovsky. 2015. Applying differential privacy to matrix factorization. In *Proceedings of the 9th ACM Conference on Recommender Systems.* 107–114.

[25] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD.*

[26] Smriti Bhagat, Irina Rozenbaum, and Graham Cormode. 2007. Applying link-based classification to label blogs. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis.* 92–101.

[27] Runa Bhaumik, Bamshad Mobasher, and Robin Burke. 2011. A clustering approach to unsupervised attack detection in collaborative recommender systems. In *Proceedings of the International Conference on Data Science (ICDATA).* Citeseer, 1.

[28] Runa Bhaumik, Chad Williams, Bamshad Mobasher, and Robin Burke. 2006. Securing collaborative filtering against malicious attacks through anomaly detection. In *Proceedings of the 4th workshop on intelligent techniques for web personalization (ITWP'06), Boston*, Vol. 6. 10.

[29] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 2013. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases.* Springer, 387–402.

[30] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender systems survey. *Knowledge-based systems* 46 (2013), 109–132.

[31] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[32] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 1175–1191.

[33] Ludovico Boratto, Gianni Fenu, and Mirko Marras. 2021. Interplay between upsampling and regularization for provider fairness in recommender systems. *User Modeling and User-Adapted Interaction* 31, 3 (2021), 421–455.

[34] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. TasteWeights: a visual interactive hybrid recommender system. In *Proceedings of the sixth ACM conference on Recommender systems*. 35–42.

[35] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[36] Robin Burke. 2002. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction* 12, 4 (2002), 331–370.

[37] Robin Burke, Bamshad Mobasher, Runa Bhaumik, and Chad Williams. 2005. Segment-based injection attacks against collaborative filtering recommender systems. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*. IEEE, 4–pp.

[38] Robin Burke, Bamshad Mobasher, Chad Williams, and Runa Bhaumik. 2006. Detecting profile injection attacks in collaborative recommender systems. In *The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE'06)*. IEEE, 23–23.

[39] Hongyun Cai and Fuzhi Zhang. 2019. An unsupervised method for detecting shilling attacks in recommender systems by mining item relationship and identifying target items. *Comput. J.* 62, 4 (2019), 579–597.

[40] ZeFeng Cai, Linlin Wang, Gerard de Melo, Fei Sun, and Liang He. 2022. Multi-Scale Distribution Deep Variational Autoencoder for Explanation Generation. In *Findings of the Association for Computational Linguistics: ACL 2022*. 68–78.

[41] Joseph A Calandrino, Ann Kilzer, Arvind Narayanan, Edward W Felten, and Vitaly Shmatikov. 2011. " You might also like:" Privacy risks of collaborative filtering. In *2011 IEEE symposium on security and privacy*. IEEE, 231–246.

[42] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 ICDM Workshops*. IEEE, 13–18.

[43] John Canny. 2002. Collaborative filtering with privacy. In *Proceedings 2002 IEEE Symposium on Security and Privacy*. IEEE, 45–57.

[44] Jie Cao, Zhiang Wu, Bo Mao, and Yanchun Zhang. 2013. Shilling attack detection utilizing semi-supervised learning method for collaborative recommender system. *World Wide Web* 16, 5 (2013), 729–748.

[45] Yuanjiang Cao, Xiaocong Chen, Lina Yao, Xianzhi Wang, and Wei Emma Zhang. 2020. Adversarial attacks and detection on reinforcement learning-based interactive recommender systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1669–1672.

[46] L Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. 2019. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the conference on fairness, accountability, and transparency*. 160–169.

[47] Abdelberi Chaabane, Gergely Acs, Mohamed Ali Kaafar, et al. 2012. You are what you like! information leakage through users' interests. In *Proceedings of the 19th annual network & distributed system security symposium (NDSS)*. Citeseer.

[48] Di Chai, Leye Wang, Kai Chen, and Qiang Yang. 2020. Secure federated matrix factorization. *IEEE Intelligent Systems* 36, 5 (2020), 11–20.

[49] Hanxiong Chen, Xu Chen, Shaoyun Shi, and Yongfeng Zhang. 2019. Generate natural language explanations for recommendation. In *Proceedings of the SIGIR 2019 Workshop on ExplainAble Recommendation and Search*.

[50] Huiyuan Chen and Jing Li. 2019. Adversarial tensor factorization for context-aware recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 363–367.

[51] Huiyuan Chen and Jing Li. 2019. Data poisoning attacks on cross-domain recommendation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2177–2180.

[52] Hanxiong Chen, Yunqi Li, Shaoyun Shi, Shuchang Liu, He Zhu, and Yongfeng Zhang. 2022. Graph collaborative reasoning. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 75–84.

[53] Hanxiong Chen, Shaoyun Shi, Yunqi Li, and Yongfeng Zhang. 2021. Neural collaborative reasoning. In *Proceedings of the Web Conference 2021*. 1516–1527.

[54] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2020. Bias and debias in recommender system: A survey and future directions. *arXiv preprint arXiv:2010.03240* (2020).

[55] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. [n.d.]. Investigating the Impact of Gender on Rank in Resume Search Engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

[56] Tong Chen, Hongzhi Yin, Guanhua Ye, Zi Huang, Yang Wang, and Meng Wang. 2020. Try this instead: Personalized and interpretable substitute recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 891–900.

[57] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 765–774.

[58] Xiaoqiang Chen and Vincent Huang. 2012. Privacy preserving data publishing for recommender system. In *2012 IEEE 36th Annual Computer Software and Applications Conference Workshops*. IEEE, 128–133.

[59] Xu Chen, Hongteng Xu, Yongfeng Zhang, Jiaxi Tang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2018. Sequential recommendation with user memory networks. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 108–116.

[60] Xu Chen, Yongfeng Zhang, and Zheng Qin. 2019. Dynamic explainable recommendation based on neural attentive models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 53–60.

[61] Xu Chen, Yongfeng Zhang, and Ji-Rong Wen. 2022. Measuring" Why" in Recommender Systems: a Comprehensive Survey on the Evaluation of Explainable Recommendation. *arXiv preprint arXiv:2202.06466* (2022).

[62] Zhongxia Chen, Xiting Wang, Xing Xie, Mehul Parsana, Akshay Soni, Xiang Ao, and Enhong Chen. 2020. Towards Explainable Conversational Recommendation.. In *IJCAI*. 2994–3000.

[63] Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. 2019. Co-Attentive Multi-Task Learning for Explainable Recommendation. In *IJCAI*. 2137–2143.

[64] Zhili Chen, Yu Wang, Shun Zhang, Hong Zhong, and Lin Chen. 2021. Differentially private user-based collaborative filtering recommendation based on K-means clustering. *Expert Systems with Applications* 168 (2021), 114366.

[65] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. 7–10.

[66] Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose C Kanjirathinkal, and Mohan Kankanhalli. 2019. MMALFM: Explainable recommendation by leveraging reviews and images. *ACM Transactions on Information Systems (TOIS)* 37, 2 (2019), 1–28.

[67] Zunping Cheng and Neil Hurley. 2009. Effective diverse and obfuscated attacks on model-based recommender systems. In *Proceedings of the third ACM conference on Recommender systems*. 141–148.

[68] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[69] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

[70] Richard Cissée and Sahin Albayrak. 2007. An agent-based approach for privacy-preserving recommender systems. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*. 1–8.

[71] Roberto Confalonieri, Ludovik Coba, Benedikt Wagner, and Tarek R Besold. 2021. A historical perspective of explainable Artificial Intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11, 1 (2021), e1391.

[72] Margaret S Crocco, Avner Segall, Anne-Lise Halvorsen, Alexandra Stamm, and Rebecca Jacobsen. 2020. "It's not like they're selling your data to dangerous people": Internet privacy, teens, and (non-) controversial public issues. *The Journal of Social Studies Research* 44, 1 (2020), 21–33.

[73] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. 2021. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems (TOIS)* 39, 2 (2021), 1–49.

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

[75] Gabriel de Souza Pereira Moreira, Sara Rabhi, Jeong Min Lee, Ronay Ak, and Even Oldridge. 2021. Transformers4Rec: Bridging the Gap between NLP and Sequential/Session-Based Recommendation. In *Fifteenth ACM Conference on Recommender Systems*. 143–153.

[76] Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2020. Adversarial machine learning in recommender systems (aml-recsys). In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 869–872.

[77] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2022. A Survey of Research on Fair Recommender Systems. *arXiv preprint arXiv:2205.11127* (2022).

[78] Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2021. A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–38.

[79] Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2021. A Survey on Adversarial Recommender Systems: From Attack/Defense Strategies to Generative Adversarial Networks. *ACM Comput. Surv.* 54, 2, Article 35 (mar 2021), 38 pages. https://doi.org/10.1145/3439729

[80] Shuiguang Deng, Hailiang Zhao, Weijia Fang, Jianwei Yin, Schahram Dustdar, and Albert Y Zomaya. 2020. Edge intelligence: The confluence of edge computing and artificial intelligence. *IEEE Internet of Things Journal* 7, 8 (2020), 7457–7469.

[81] Ratan Dey, Cong Tang, Keith Ross, and Nitesh Saxena. 2012. Estimating age privacy leakage in online social networks. In *2012 proceedings ieee infocom*. IEEE, 2836–2840.

[82] Manqing Dong, Feng Yuan, Lina Yao, Xianzhi Wang, Xiwei Xu, and Liming Zhu. 2020. Survey for Trust-aware Recommender Systems: A Deep Learning Perspective. *arXiv preprint arXiv:2004.03774* (2020).

[83] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.

[84] Yali Du, Meng Fang, Jinfeng Yi, Chang Xu, Jun Cheng, and Dacheng Tao. 2018. Enhancing the robustness of neural collaborative filtering systems under malicious attacks. *IEEE Transactions on Multimedia* 21, 3 (2018), 555–565.

[85] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[86] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. 2016. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853* (2016).

[87] Zekeriya Erkin, Thijs Veugen, Tomas Toft, and Reginald L Lagendijk. 2012. Generating private recommendations efficiently using homomorphic encryption and data packing. *IEEE transactions on information forensics and security* 7, 3 (2012), 1053–1066.

[88] Hui Fang, Danning Zhang, Yiheng Shu, and Guibing Guo. 2020. Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations. *ACM Transactions on Information Systems (TOIS)* 39, 1 (2020), 1–42.

[89] Minghong Fang, Neil Zhenqiang Gong, and Jia Liu. 2020. Influence function based data poisoning attacks to top-n recommender systems. In *Proceedings of The Web Conference 2020*. 3019–3025.

[90] Minghong Fang, Guolei Yang, Neil Zhenqiang Gong, and Jia Liu. 2018. Poisoning attacks to graph-based recommender systems. In *Proceedings of the 34th Annual Computer Security Applications Conference*. 381–392.

[91] Maurizio Ferrari Dacrema, Federico Parroni, Paolo Cremonesi, and Dietmar Jannach. 2020. Critically examining the claimed value of convolutions over user-item embedding maps for recommender systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 355–363.

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

[93] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 1322–1333.

[94] Remo Frey, Dominic Wörner, and Alexander Ilic. 2016. Collaborative filtering on the blockchain: a secure recommender system for e-commerce. (2016).

[95] Arik Friedman, Bart P Knijnenburg, Kris Vanhecke, Luc Martens, and Shlomo Berkovsky. 2015. Privacy aspects of recommender systems. In *Recommender systems handbook*. Springer, 649–688.

[96] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, et al. 2020. Fairness-aware explainable recommendation over knowledge graphs. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 69–78.

[97] Zuohui Fu, Yikun Xian, Yaxin Zhu, Shuyuan Xu, Zelong Li, Gerard De Melo, and Yongfeng Zhang. 2021. HOOPS: Human-in-the-Loop Graph Reasoning for Conversational Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2415–2421.

[98] Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. 2008. Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 265–273.

[99] Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 50–56.

[100] Ruoyuan Gao, Yingqiang Ge, and Chirag Shah. 2021. FAIR: Fairness-Aware Information Retrieval Evaluation. *arXiv preprint arXiv:2106.08527* (2021).

[101] Ruoyuan Gao and Chirag Shah. 2021. Addressing Bias and Fairness in Search Systems. In *Proceedings of the 44th International ACM SIGIR (SIGIR '21)*. 4. https://doi.org/10.1145/3404835.3462807

[102] Yang Gao, Yi-Fan Li, Yu Lin, Hang Gao, and Latifur Khan. 2020. Deep learning on knowledge graph for recommender system: A survey. *arXiv preprint arXiv:2004.00387* (2020).

[103] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, and Yongfeng Zhang. 2021. Towards Long-term Fairness in Recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 445–453.

[104] Yingqiang Ge, Juntao Tan, Yan Zhu, Yinglong Xia, Jiebo Luo, Shuchang Liu, Zuohui Fu, Shijie Geng, Zelong Li, and Yongfeng Zhang. 2022. Explainable Fairness in Recommendation. *SIGIR* (2022).

[105] Yingqiang Ge, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. 2022. Toward Pareto Efficient Fairness-Utility Trade-off in Recommendation through Reinforcement Learning. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*.

[106] Shijie Geng, Zuohui Fu, Yingqiang Ge, Lei Li, Gerard de Melo, and Yongfeng Zhang. 2022. Improving Personalized Explanation Generation through Visualization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 244–255.

[107] Shijie Geng, Zuohui Fu, Juntao Tan, Yingqiang Ge, Gerard De Melo, and Yongfeng Zhang. 2022. Path Language Modeling over Knowledge Graphsfor Explainable Recommendation. In *Proceedings of the ACM Web Conference 2022*. 946–955.

[108] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). In *Proceedings of the 16th ACM Conference on Recommender Systems*.

[109] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. In *Proceedings of KDD*. ACM, 2221–2231.

[110] Alireza Gharahighehi, Celine Vens, and Konstantinos Pliakos. 2021. Fair multi-stakeholder news recommender system with hypergraph ranking. *Information Processing & Management* 58, 5 (2021), 102663.

[111] Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy, and Gerhard Weikum. 2020. PRINCE: Provider-side interpretability with counterfactual explanations in recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 196–204.

[112] Boris Glavic, Alexandra Meliou, and Sudeepa Roy. 2021. Trends in explanations: Understanding and debugging data-driven systems. *Foundations and Trends® in Databases* 11, 3 (2021).

[113] Elizabeth Gómez, Carlos Shui Zhang, Ludovico Boratto, Maria Salamó, and Mirko Marras. 2021. 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*. 1808–1812.

[114] Neil Zhenqiang Gong and Bin Liu. 2016. You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors. In *25th USENIX Security Symposium (USENIX Security 16)*. 979–995.

[115] Yu Gong, Ziwen Jiang, Yufei Feng, Binbin Hu, Kaiqi Zhao, Qingwen Liu, and Wenwu Ou. 2020. EdgeRec: recommender system on edge in Mobile Taobao. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2477–2484.

[116] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[117] Shixiang Gu and Luca Rigazio. 2014. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068* (2014).

[118] Ihsan Gunes, Cihan Kaleli, Alper Bilge, and Huseyin Polat. 2014. Shilling attacks against recommender systems: a comprehensive survey. *Artificial Intelligence Review* 42, 4 (2014), 767–799.

[119] David Gunning and David Aha. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI magazine* 40, 2 (2019), 44–58.

[120] Deepesh V Hada and Shirish K Shevade. 2021. ReXPlug: Explainable Recommendation using Plug-and-Play Language Model. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 81–91.

[121] Kishaloy Halder, Min-Yen Kan, and Kazunari Sugiyama. 2017. Health forum thread recommendation using an interest aware topic model. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1589–1598.

[122] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *NeurIPS*. 3315–3323.

[123] F. Maxwell Harper, Funing Xu, Harmanpreet Kaur, Kyle Condiff, Shuo Chang, and Loren Terveen. 2015. Putting Users in Control of Their Recommendations *(RecSys '15)*. Association for Computing Machinery, New York, NY, USA, 3–10. https://doi.org/10.1145/2792838.2800179

[124] Jianming He, Wesley W Chu, and Zhenyu Victor Liu. 2006. Inferring privacy information from social networks. In *International Conference on Intelligence and Security Informatics*. Springer, 154–165.
[125] Ruining He and Julian McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *AAAI*.
[126] Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. 2018. Adversarial Personalized Ranking for Recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 355–364.
[127] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *WWW*. 173–182.
[128] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)* 22, 1 (2004), 5–53.
[129] Seira Hidano and Shinsaku Kiyomoto. 2020. Recommender Systems Robust to Data Poisoning using Trim Learning.. In *ICISSP*. 721–724.
[130] Seira Hidano, Takao Murakami, Shuichi Katsumata, Shinsaku Kiyomoto, and Goichiro Hanaoka. 2017. Model inversion attacks for prediction systems: Without knowledge of non-sensitive attributes. In *2017 15th Annual Conference on Privacy, Security and Trust (PST)*. IEEE, 115–11509.
[131] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).
[132] Yoshinori Hijikata, Kazuhiro Iwahama, and Shogo Nishida. 2006. Content-based music filtering system with editable user profile. In *Proceedings of the 2006 ACM symposium on Applied computing*. 1050–1057.
[133] Yoshinori Hijikata, Yuki Kai, and Shogo Nishida. 2012. The relation between user intervention and user satisfaction for information recommendation. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*. 2002–2007.
[134] Yassine Himeur, Abdullah Alsalemi, Ayman Al-Kababji, Faycal Bensaali, Abbes Amira, Christos Sardianos, George Dimitrakopoulos, and Iraklis Varlamis. 2021. A survey of recommender systems for energy efficiency in buildings: Principles, challenges and prospects. *Information Fusion* 72 (2021), 1–21.
[135] Yassine Himeur, Khalida Ghanem, Abdullah Alsalemi, Faycal Bensaali, and Abbes Amira. 2021. Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. *Applied Energy* 287 (2021), 116601.
[136] Yassine Himeur, Shahab Saquib Sohail, Faycal Bensaali, Abbes Amira, and Mamoun Alazab. 2022. Latest Trends of Security and Privacy in Recommender Systems: A Comprehensive Review and Future Perspectives. *Computers & Security* (2022), 102746.
[137] Cheng-Kang Hsieh, Longqi Yang, Yin Cui, Tsung-Yi Lin, Serge Belongie, and Deborah Estrin. 2017. Collaborative metric learning. In *Proceedings of the 26th international conference on world wide web*. 193–201.
[138] Haoji Hu, Xiangnan He, Jinyang Gao, and Zhi-Li Zhang. 2020. 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*. 1071–1080.
[139] Jin Huang, Wayne Xin Zhao, Hongjian Dou, Ji-Rong Wen, and Edward Y Chang. 2018. Improving sequential recommendation with knowledge-enhanced memory networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 505–514.
[140] Weiming Huang, Baisong Liu, and Hao Tang. 2019. Privacy protection for recommendation system: a survey. In *Journal of Physics: Conference Series*, Vol. 1325. IOP Publishing, 012087.
[141] Dietmar Jannach and Michael Jugovac. 2019. Measuring the Business Value of Recommender Systems. *ACM Trans. Manage. Inf. Syst.* 10, 4, Article 16 (dec 2019), 23 pages. https://doi.org/10.1145/3370082
[142] Dietmar Jannach, Michael Jugovac, and Ingrid Nunes. 2019. Explanations and user control in recommender systems. In *Proceedings of the 23rd International Workshop on Personalization and Recommendation on the Web and Beyond*. 31–31.
[143] Dietmar Jannach, Sidra Naveed, and Michael Jugovac. 2016. User control in recommender systems: Overview and interaction challenges. In *International Conference on Electronic Commerce and Web Technologies*. Springer, 21–33.
[144] Jinyuan Jia and Neil Zhenqiang Gong. 2018. {AttriGuard}: A Practical Defense Against Attribute Inference Attacks via Adversarial Machine Learning. In *27th USENIX Security Symposium (USENIX Security 18)*. 513–529.
[145] Jia-Yun Jiang, Cheng-Te Li, and Shou-De Lin. 2019. Towards a more reliable privacy-preserving recommender system. *Information Sciences* 482 (2019), 248–265.
[146] Michael Jugovac and Dietmar Jannach. 2017. Interacting with recommenders—overview and research directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 3 (2017), 1–46.
[147] Vassilis Kaffes, Dimitris Sacharidis, and Giorgos Giannopoulos. 2021. Model-Agnostic Counterfactual Explanations of Recommendations. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 280–285.
[148] Jerry Kang. 1997. Information privacy in cyberspace transactions. *Stan. L. Rev.* 50 (1997), 1193.

[149] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.

[150] A. Khademi, S. Lee, D. Foley, and V. Honavar. 2019. Fairness in algorithmic decision making: An excursion through the lens of causality. In *WWW*.

[151] Ömer Kırnap, Fernando Diaz, Asia Biega, Michael Ekstrand, Ben Carterette, and Emine Yilmaz. 2021. Estimation of Fair Ranking Metrics with Incomplete Judgments. In *Proceedings of the Web Conference 2021*. 1065–1075.

[152] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, Vol. 67. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 43.

[153] Bart P Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems*. 43–50.

[154] Bart P Knijnenburg, Niels JM Reijmer, and Martijn C Willemsen. 2011. Each to his own: how different users call for different interaction methods in recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. 141–148.

[155] Michal Kompan and Mária Bieliková. 2010. Content-based news recommendation. In *International conference on electronic commerce and web technologies*. Springer, 61–72.

[156] Joseph A Konstan, Bradley N Miller, David Maltz, Jonathan L Herlocker, Lee R Gordon, and John Riedl. 1997. GroupLens: applying collaborative filtering to Usenet news. *Commun. ACM* 40, 3 (1997), 77–87.

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

[158] Balachander Krishnamurthy and Craig E Wills. 2009. On the leakage of personally identifiable information via online social networks. In *Proceedings of the 2nd ACM workshop on Online social networks*. 7–12.

[159] Aditya Kuppa and Nhien-An Le-Khac. 2020. Black box attacks on explainable artificial intelligence (XAI) methods in cyber security. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[160] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *NeurIPS*. 4069–4079.

[161] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. 2019. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 ieee 35th international conference on data engineering (icde)*. IEEE, 1334–1345.

[162] Shyong K Lam and John Riedl. 2004. Shilling recommender systems for fun and profit. In *Proceedings of the 13th international conference on World Wide Web*. 393–402.

[163] Shyong K. Lam and John Riedl. 2004. Shilling Recommender Systems for Fun and Profit. In *Proceedings of the 13th International Conference on World Wide Web (WWW '04)*. Association for Computing Machinery, New York, NY, USA, 393–402. https://doi.org/10.1145/988672.988726

[164] Béatrice Lamche, Ugur Adıgüzel, and Wolfgang Wörndl. 2014. Interactive explanations in mobile shopping recommender systems. In *Joint Workshop on Interfaces and Human Decision Making in Recommender Systems*, Vol. 14.

[165] Yixiao Lan, Yuan Liu, Boyang Li, and Chunyan Miao. 2021. Proof of Learning (PoLe): Empowering machine learning with consensus building on blockchains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 16063–16066.

[166] Trung-Hoang Le and Hady W Lauw. 2021. Explainable recommendation with comparative constraints on product aspects. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 967–975.

[167] Bo Li, Yining Wang, Aarti Singh, and Yevgeniy Vorobeychik. 2016. Data poisoning attacks on factorization-based collaborative filtering. *Advances in neural information processing systems* 29 (2016).

[168] Huaxin Li, Qingrong Chen, Haojin Zhu, Di Ma, Hong Wen, and Xuemin Sherman Shen. 2017. Privacy leakage via de-anonymization and aggregation in heterogeneous social networks. *IEEE Transactions on Dependable and Secure Computing* 17, 2 (2017), 350–362.

[169] Lei Li, Li Chen, and Yongfeng Zhang. 2020. Towards controllable explanation generation for recommender systems via neural template. In *Companion proceedings of the web conference 2020*. 198–202.

[170] Lei Li, Yongfeng Zhang, and Li Chen. 2020. Generate neural template explanations for recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 755–764.

[171] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Extra: Explanation ranking datasets for explainable recommendation. In *Proceedings of the 44th International ACM SIGIR conference on Research and Development in Information Retrieval*. 2463–2469.

[172] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized Transformer for Explainable Recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4947–4957.

[173] Lei Li, Yongfeng Zhang, and Li Chen. 2022. Personalized Prompt Learning for Explainable Recommendation. *arXiv preprint arXiv:2202.07371* (2022).

[174] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering*. IEEE, 106–115.
[175] Ting Li and Till Unger. 2012. Willing to pay for quality personalization? Trade-off between quality and privacy. *European Journal of Information Systems* 21, 6 (2012), 621–642.
[176] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented Fairness in Recommendation. In *Proceedings of the Web Conference 2021*. 624–632.
[177] Yunqi Li, Hanxiong Chen, Juntao Tan, and Yongfeng Zhang. 2022. Causal factorization machine for robust recommendation. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*. 1–9.
[178] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. 2022. Fairness in Recommendation: A Survey. *arXiv preprint arXiv:2205.13619* (2022).
[179] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2021. Towards personalized fairness based on causal notion. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1054–1063.
[180] Yunqi Li, Yingqiang Ge, and Yongfeng Zhang. 2021. Tutorial on Fairness of Machine Learning in Recommender Systems. In *Proceedings of the 30th CIKM*.
[181] Zelong Li, Jianchao Ji, and Yongfeng Zhang. 2022. From Kepler to Newton: Explainable AI for Science Discovery. In *ICML 2022 2nd AI for Science Workshop*. https://openreview.net/forum?id=vA9hti-Fi7H
[182] Xiao Lin, Min Zhang, Yongfeng Zhang, Zhaoquan Gu, Yiqun Liu, and Shaoping Ma. 2017. Fairness-aware group recommendation with pareto-efficiency. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 107–115.
[183] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon.Com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing* 7, 1 (Jan. 2003), 76–80. https://doi.org/10.1109/MIC.2003.1167344
[184] Andrea Lisi, Andrea De Salve, Paolo Mori, and Laura Ricci. 2020. Practical application and evaluation of atomic swaps for blockchain-based recommender systems. In *2020 the 3rd International Conference on Blockchain Technology and Applications*. 67–74.
[185] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2021. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–36.
[186] Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Yunhao Liu, Anil K Jain, and Jiliang Tang. 2021. Trustworthy ai: A computational perspective. *arXiv preprint arXiv:2107.06641* (2021).
[187] Huafeng Liu, Jingxuan Wen, Liping Jing, Jian Yu, Xiangliang Zhang, and Min Zhang. 2019. In2Rec: Influence-based interpretable recommendation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1803–1812.
[188] Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. 2018. STAMP: short-term attention/memory priority model for session-based recommendation. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1831–1839.
[189] Shuchang Liu, Yingqiang Ge, Shuyuan Xu, Yongfeng Zhang, and Amelie Marian. 2022. Fairness-aware Federated Matrix Factorization. In *RecSys*.
[190] Yang Liu, Xianzhuo Xia, Liang Chen, Xiangnan He, Carl Yang, and Zibin Zheng. 2020. Certifiable robustness to discrete adversarial perturbations for factorization machines. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 419–428.
[191] Zhuoran Liu and Martha Larson. 2021. Adversarial Item Promotion: Vulnerabilities at the Core of Top-N Recommenders that Use Images to Address Cold Start. In *Proceedings of the Web Conference 2021*. 3590–3602.
[192] Yichao Lu, Ruihai Dong, and Barry Smyth. 2018. Why I like it: multi-task learning for recommendation and explanation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 4–12.
[193] Chen Ma, Peng Kang, and Xue Liu. 2019. Hierarchical gating networks for sequential recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 825–833.
[194] Weizhi Ma, Min Zhang, Yue Cao, Woojeong Jin, Chenyang Wang, Yiqun Liu, Shaoping Ma, and Xiang Ren. 2019. Jointly learning explainable rules for recommendation with knowledge graph. In *The world wide web conference*. 1210–1221.
[195] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 3–es.
[196] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In *SIGIR*. ACM.
[197] Sean M McNee, Shyong K Lam, Joseph A Konstan, and John Riedl. 2003. Interfaces for eliciting new user preferences in recommender systems. In *International Conference on User Modeling*. Springer, 178–187.

[198] Sean M McNee, John Riedl, and Joseph A Konstan. 2006. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*. 1097–1101.

[199] Sean M McNee, John Riedl, and Joseph A Konstan. 2006. Making recommendations better: an analytic model for human-recommender interaction. In *CHI'06 extended abstracts on Human factors in computing systems*. 1103–1108.

[200] Frank McSherry and Ilya Mironov. 2009. Differentially private recommender systems: Building privacy into the netflix prize contenders. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 627–636.

[201] Shagufta Mehnaz, Sayanton V Dibbo, Ehsanul Kabir, Ninghui Li, and Elisa Bertino. 2022. Are Your Sensitive Attributes Private? Novel Model Inversion Attribute Inference Attacks on Classification Models. (Aug. 2022).

[202] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[203] Bhaskar Mehta. 2007. Unsupervised shilling detection for collaborative filtering. In *AAAI*. 1402–1407.

[204] Bhaskar Mehta and Wolfgang Nejdl. 2009. Unsupervised strategies for shilling detection and robust collaborative filtering. *User Modeling and User-Adapted Interaction* 19, 1 (2009), 65–97.

[205] Christian Meske, Enrico Bunde, Johannes Schneider, and Martin Gersch. 2022. Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Information Systems Management* 39, 1 (2022), 53–63.

[206] Andriy Mnih and Ruslan R Salakhutdinov. 2008. Probabilistic matrix factorization. In *Advances in neural information processing systems*. 1257–1264.

[207] Bamshad Mobasher, Robin Burke, Runa Bhaumik, and Chad Williams. 2007. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Transactions on Internet Technology (TOIT)* 7, 4 (2007), 23–es.

[208] Christoph Molnar. 2020. *Interpretable machine learning*. Lulu. com.

[209] Khalil Muhammad, Qinqin Wang, Diarmuid O'Reilly-Morgan, Elias Tragos, Barry Smyth, Neil Hurley, James Geraci, and Aonghus Lawlor. 2020. Fedfast: Going beyond average for faster training of federated recommender systems. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1234–1242.

[210] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 111–125.

[211] Valeria Nikolaenko, Stratis Ioannidis, Udi Weinsberg, Marc Joye, Nina Taft, and Dan Boneh. 2013. Privacy-preserving matrix factorization. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. 801–812.

[212] Paul Ohm. 2009. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA l. Rev.* 57 (2009), 1701.

[213] Michael P O'mahony, Neil J Hurley, and Guenole CM Silvestre. 2004. An evaluation of neighbourhood formation on the performance of collaborative filtering. *Artificial Intelligence Review* 21, 3 (2004), 215–228.

[214] Michael P O'Mahony, Neil J Hurley, and Guénolé CM Silvestre. 2005. Recommender systems: Attack types and strategies. In *AAAI*. 334–339.

[215] Behrooz Omidvar-Tehrani, Sruthi Viswanathan, and Jean-Michel Renders. 2020. Interactive and explainable point-of-interest recommendation using look-alike groups. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*. 389–392.

[216] Jeroen Ooge, Shotallo Kato, and Katrien Verbert. 2022. Explaining Recommendations in E-Learning: Effects on Adolescents' Trust. In *27th International Conference on Intelligent User Interfaces*. 93–105.

[217] Mourad Oussalah. 2021. AI Explainability. A Bridge Between Machine Vision and Natural Language Processing. In *International Conference on Pattern Recognition*. Springer, 257–273.

[218] Michael P O'Mahony, Neil J Hurley, and Guenole Silvestre. 2002. Promoting recommendations: An attack on collaborative filtering. In *International Conference on Database and Expert Systems Applications*. Springer, 494–503.

[219] Mirko Palato. 2021. Federated Variational Autoencoder for Collaborative Filtering. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[220] Sicheng Pan, Dongsheng Li, Hansu Gu, Tun Lu, Xufang Luo, and Ning Gu. 2022. Accurate and Explainable Recommendation via Review Rationalization. In *Proceedings of the ACM Web Conference 2022*. 3092–3101.

[221] Weishen Pan, Sen Cui, Jiang Bian, Changshui Zhang, and Fei Wang. 2021. Explaining algorithmic fairness through fairness-aware causal path decomposition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1287–1297.

[222] Stuart L Pardau. 2018. The California consumer privacy act: Towards a European-style privacy regime in the United States. *J. Tech. L. & Pol'y* 23 (2018), 68.

[223] Michael J Pazzani and Daniel Billsus. 2007. Content-based recommendation systems. In *The adaptive web*. Springer, 325–341.

[224] Georgina Peake and Jun Wang. 2018. 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*. 2060–2069.

[225] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2021. Fairness in rankings and recommendations: An overview. *The VLDB Journal* (2021), 1–28.

[226] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems* 30 (2017).

[227] Huseyin Polat and Wenliang Du. 2003. Privacy-preserving collaborative filtering using randomized perturbation techniques. In *Third IEEE International Conference on Data Mining*. IEEE, 625–628.

[228] Talya Porat, Rune Nyrup, Rafael A Calvo, Priya Paudyal, and Elizabeth Ford. 2020. Public health and risk communication during COVID-19—enhancing psychological needs to promote sustainable behavior change. *Frontiers in public health* (2020), 637.

[229] Reinald Adrian Pugoy and Hung-Yu Kao. 2020. BERT-based neural collaborative filtering and fixed-length contiguous tokens explanation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. 143–153.

[230] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[231] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 140 (2020), 1–67.

[232] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

[233] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[234] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. 37–44.

[235] Zhaochun Ren, Shangsong Liang, Piji Li, Shuaiqiang Wang, and Maarten de Rijke. 2017. Social collaborative viewpoint regression with explainable recommendations. In *Proceedings of the tenth ACM international conference on web search and data mining*. 485–494.

[236] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 452–461.

[237] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*. 811–820.

[238] Steffen Rendle, Walid Krichene, Li Zhang, and John Anderson. 2020. Neural collaborative filtering vs. matrix factorization revisited. In *Fourteenth ACM conference on recommender systems*. 240–248.

[239] Yehezkel S Resheff, Yanai Elazar, Moni Shahar, and Oren Sar Shalom. 2018. Privacy and fairness in recommender systems via adversarial training of user representations. *arXiv preprint arXiv:1807.03521* (2018).

[240] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. ACM, 175–186.

[241] Christos Sardianos, Iraklis Varlamis, Christos Chronis, George Dimitrakopoulos, Abdullah Alsalemi, Yassine Himeur, Faycal Bensaali, and Abbes Amira. 2021. The emergence of explainability of intelligent systems: Delivering explainable and personalized recommendations for energy efficiency. *International Journal of Intelligent Systems* 36, 2 (2021), 656–680.

[242] Alper Kaan Sarica and Pelin Angin. 2020. Explainable security in SDN-based IoT networks. *Sensors* 20, 24 (2020), 7326.

[243] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. ACM, 285–295.

[244] James Schaffer, Tobias Hollerer, and John O'Donovan. 2015. Hypothetical recommendation: A study of interactive profile manipulation behavior for recommender systems. In *The Twenty-Eighth International Flairs Conference*.

[245] Sven Schmit and Carlos Riquelme. 2018. Human Interaction with Recommendation Systems. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research)*, Amos Storkey and Fernando Perez-Cruz (Eds.), Vol. 84. PMLR, 862–870. https://proceedings.mlr.press/v84/schmit18a.html

[246] Dougal Shakespeare, Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. 2020. Exploring artist gender bias in music recommendation. *arXiv preprint arXiv:2009.01715* (2020).
[247] Lloyd S Shapley. 2016. *17. A value for n-person games.* Princeton University Press.
[248] Shaoyun Shi, Hanxiong Chen, Weizhi Ma, Jiaxin Mao, Min Zhang, and Yongfeng Zhang. 2020. Neural logic reasoning. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1365–1374.
[249] Yue Shi, Martha Larson, and Alan Hanjalic. 2014. Collaborative Filtering beyond the User-Item Matrix: A Survey of the State of the Art and Future Challenges. *ACM Comput. Surv.* 47, 1, Article 3 (may 2014), 45 pages. https://doi.org/10.1145/2556270
[250] Donghee Shin. 2020. How do users interact with algorithm recommender systems? The interaction of users, algorithms, and performance. *Computers in Human Behavior* 109 (2020), 106344.
[251] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 3–18.
[252] Liu Shuchang, Ge Yingqiang, Shuyuan Xu, Yongfeng Zhang, and Amélie Marian. 2022. Fairness-aware Federated Matrix Factorization. In *2022 RecSys*.
[253] Mingdan Si and Qingshan Li. 2020. Shilling attacks against collaborative recommender systems: a review. *Artificial Intelligence Review* 53, 1 (2020), 291–319.
[254] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD*.
[255] H Jeff Smith, Tamara Dinev, and Heng Xu. 2011. Information privacy research: an interdisciplinary review. *MIS quarterly* (2011), 989–1015.
[256] S Sobitha Ahila and KL Shunmuganathan. 2016. Role of agent technology in web usage mining: homomorphic encryption based recommendation for e-commerce applications. *Wireless Personal Communications* 87, 2 (2016), 499–512.
[257] Daniel J Solove. 2005. A taxonomy of privacy. *U. Pa. l. Rev.* 154 (2005), 477.
[258] Piotr Sulikowski, Tomasz Zdziebko, Dominik Turzyński, and Eliasz Kańtoch. 2018. Human-website interaction monitoring in recommender systems. *Procedia Computer Science* 126 (2018), 1587–1596.
[259] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.
[260] Peijie Sun, Le Wu, Kun Zhang, Yanjie Fu, Richang Hong, and Meng Wang. 2020. Dual learning for explainable recommendation: Towards unifying user preference prediction and review generation. In *Proceedings of The Web Conference 2020*. 837–847.
[261] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.
[262] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos. 2009. MoviExplain: a recommender system with explanations. In *Proceedings of the third ACM conference on Recommender systems*. 317–320.
[263] Chang-You Tai, Liang-Ying Huang, Chien-Kun Huang, and Lun-Wei Ku. 2021. User-Centric Path Reasoning towards Explainable Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 879–889.
[264] Kyosuke Takami, Yiling Dai, Brendan Flanagan, and Hiroaki Ogata. 2022. Educational Explainable Recommender Usage and its Effectiveness in High School Summer Vacation Assignment. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. 458–464.
[265] Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, and Yongfeng Zhang. 2022. Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning. In *Proceedings of the ACM Web Conference 2022*. 1018–1027.
[266] Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. 2021. Counterfactual explainable recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1784–1793.
[267] Jinhui Tang, Xiaoyu Du, Xiangnan He, Fajie Yuan, Qi Tian, and Tat-Seng Chua. 2019. Adversarial training towards robust multimedia recommender system. *IEEE Transactions on Knowledge and Data Engineering* 32, 5 (2019), 855–867.
[268] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 565–573.
[269] Manolis Terrovitis, John Liagouris, Nikos Mamoulis, and Spiros Skiadopoulos. 2012. Privacy Preservation by Disassociation. *Proceedings of the VLDB Endowment* 5 (06 2012). https://doi.org/10.14778/2336664.2336668
[270] Loren Terveen, Jessica McMackin, Brian Amento, and Will Hill. 2002. Specifying preferences based on user history. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 315–322.
[271] Marko Tesic. 2020. *Explanation and argument.* Ph.D. Dissertation. Birkbeck, University of London.

[272] Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender systems handbook*. Springer, 353–382.

[273] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. 2020. The Relationship between Trust in AI and Trustworthy Machine Learning Technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 272–283. https://doi.org/10.1145/3351095.3372834

[274] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing Machine Learning Models via Prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*. 601–618.

[275] Khanh Hiep Tran, Azin Ghazimatin, and Rishiraj Saha Roy. 2021. Counterfactual Explanations for Neural Recommenders. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1627–1631.

[276] Chun-Hua Tsai. 2019. *Controllability and explainability in a hybrid social recommender system.* Ph.D. Dissertation. University of Pittsburgh.

[277] Kazutoshi Umemoto, Tova Milo, and Masaru Kitsuregawa. 2020. Toward recommendation for upskilling: Modeling skill improvement and item difficulty in action sequences. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 169–180.

[278] Robin Van Meteren and Maarten Van Someren. 2000. Using content-based filtering for recommendation. In *Proceedings of the machine learning in the new information age: MLnet/ECML2000 workshop*, Vol. 30. 47–56.

[279] Kush R Varshney. 2019. Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 26–29.

[280] Kush R Varshney. 2022. Trustworthy Machine Learning. *Chappaqua, NY, USA: Independently Published* (2022).

[281] Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10, 3152676 (2017), 10–5555.

[282] Cong Wang, Yifeng Zheng, Jinghua Jiang, and Kui Ren. 2018. Toward privacy-preserving personalized recommendation services. *Engineering* 4, 1 (2018), 21–28.

[283] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable recommendation via multi-task learning in opinionated text data. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 165–174.

[284] Shoujin Wang, Liang Hu, Longbing Cao, Xiaoshui Huang, Defu Lian, and Wei Liu. 2018. Attention-based transactional context embedding for next-item recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[285] Xiting Wang, Yiru Chen, Jie Yang, Le Wu, Zhengtao Wu, and Xing Xie. 2018. A reinforcement learning framework for explainable recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 587–596.

[286] Xiang Wang, Xiangnan He, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. 2018. Tem: Tree-enhanced embedding model for explainable recommendation. In *Proceedings of the 2018 World Wide Web Conference*. 1543–1552.

[287] Xiang Wang, Dingxian Wang, Canran Xu, Xiangnan He, Yixin Cao, and Tat-Seng Chua. 2019. Explainable reasoning over knowledge graphs for recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5329–5336.

[288] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2022. A Survey on the Fairness of Recommender Systems. *arXiv preprint arXiv:2206.03761* (2022).

[289] Rainer Wasinger, James Wallbank, Luiz Pizzato, Judy Kay, Bob Kummerfeld, Matthias Böhmer, and Antonio Krüger. 2013. Scrutable user models and personalised item recommendation in mobile lifestyle applications. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 77–88.

[290] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent Abilities of Large Language Models. *arXiv preprint arXiv:2206.07682* (2022).

[291] Gerhard Weikum, Xin Luna Dong, Simon Razniewski, Fabian Suchanek, et al. 2021. Machine knowledge: Creation and curation of comprehensive knowledge bases. *Foundations and Trends® in Databases* 10, 2-4 (2021), 108–490.

[292] Udi Weinsberg, Smriti Bhagat, Stratis Ioannidis, and Nina Taft. 2012. s. In *Proceedings of the sixth ACM conference on Recommender systems*. 195–202.

[293] Chad Williams and Bamshad Mobasher. 2006. Profile injection attack detection for securing collaborative recommender systems. *DePaul University CTI Technical Report* (2006), 1–47.

[294] Chad A Williams, Bamshad Mobasher, and Robin Burke. 2007. Defending recommender systems: detection of profile injection attacks. *Service Oriented Computing and Applications* 1, 3 (2007), 157–170.

[295] Chuhan Wu, Fangzhao Wu, Yang Cao, Yongfeng Huang, and Xing Xie. 2021. Fedgnn: Federated graph neural network for privacy-preserving recommendation. *arXiv preprint arXiv:2102.04925* (2021).

[296] Chao-Yuan Wu, Amr Ahmed, Alex Beutel, Alexander J Smola, and How Jing. 2017. Recurrent recommender networks. In *Proceedings of the tenth ACM international conference on web search and data mining*. 495–503.

[297] Junda Wu, Tong Yu, and Shuai Li. 2021. Deconfounded and Explainable Interactive Vision-Language Retrieval of Complex Scenes. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2103–2111.

[298] Yiqing Wu, Ruobing Xie, Yongchun Zhu, Fuzhen Zhuang, Xiang Ao, Xu Zhang, Leyu Lin, and Qing He. 2022. Selective Fairness in Recommendation via Prompts. *arXiv preprint arXiv:2205.04682* (2022).

[299] Yikun Xian. 2021. *Neural Graph Reasoning for Explainable Decision-Making*. Ph.D. Dissertation. Rutgers The State University of New Jersey, School of Graduate Studies.

[300] Yikun Xian, Zuohui Fu, Qiaoying Huang, Shan Muthukrishnan, and Yongfeng Zhang. 2020. Neural-symbolic reasoning over knowledge graph for multi-stage explainable recommendation. *arXiv preprint arXiv:2007.13207* (2020).

[301] Yikun Xian, Zuohui Fu, Shan Muthukrishnan, Gerard De Melo, and Yongfeng Zhang. 2019. Reinforcement knowledge graph reasoning for explainable recommendation. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*. 285–294.

[302] Yikun Xian, Zuohui Fu, Handong Zhao, Yingqiang Ge, Xu Chen, Qiaoying Huang, Shijie Geng, Zhou Qin, Gerard De Melo, Shan Muthukrishnan, and Yongfeng Zhang. 2020. CAFE: Coarse-to-fine neural symbolic reasoning for explainable recommendation. In *CIKM*.

[303] Yikun Xian, Handong Zhao, Tak Yeon Lee, Sungchul Kim, Ryan Rossi, Zuohui Fu, Gerard De Melo, and Shan Muthukrishnan. 2021. Exacta: Explainable column annotation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3775–3785.

[304] Yikun Xian, Tong Zhao, Jin Li, Jim Chan, Andrey Kan, Jun Ma, Xin Luna Dong, Christos Faloutsos, George Karypis, Shan Muthukrishnan, and Yongfeng Zhang. 2021. Ex3: Explainable attribute-aware item-set recommendations. In *Fifteenth ACM Conference on Recommender Systems*. 484–494.

[305] Bo Xiao and Izak Benbasat. 2007. E-commerce product recommendation agents: Use, characteristics, and impact. *MIS quarterly* (2007), 137–209.

[306] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. 2017. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*. 1369–1378.

[307] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing*. Springer, 563–574.

[308] Kun Xu and Zheng Yan. 2016. Privacy protection in mobile recommender systems: A survey. In *International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage*. Springer, 305–318.

[309] Shuyuan Xu, Yingqiang Ge, Yunqi Li, Zuohui Fu, Xu Chen, and Yongfeng Zhang. 2021. Causal collaborative filtering. *arXiv preprint arXiv:2102.01868* (2021).

[310] Shuyuan Xu, Yunqi Li, Shuchang Liu, Zuohui Fu, Yingqiang Ge, Xu Chen, and Yongfeng Zhang. 2021. Learning causal explanations for recommendation. In *The 1st International Workshop on Causality in Search and Recommendation*.

[311] Shuyuan Xu, Juntao Tan, Shelby Heinecke, Jia Li, and Yongfeng Zhang. 2021. Deconfounded Causal Collaborative Filtering. *arXiv preprint arXiv:2110.07122* (2021).

[312] Hong-Jian Xue, Xinyu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2017. Deep matrix factorization models for recommender systems.. In *IJCAI*, Vol. 17. Melbourne, Australia, 3203–3209.

[313] Roman Yampolskiy. 2020. *On controllability of artificial intelligence*. Technical Report.

[314] Aobo Yang, Nan Wang, Hongbo Deng, and Hongning Wang. 2021. Explanation as a Defense of Recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 1029–1037.

[315] Guolei Yang, Neil Zhenqiang Gong, and Ying Cai. 2017. Fake Co-visitation Injection Attacks to Recommender Systems.. In *NDSS*.

[316] Ning Yang, Yuchi Ma, Li Chen, and Philip S Yu. 2020. A meta-feature based unified framework for both cold-start and warm-start explainable recommendations. *World Wide Web* 23, 1 (2020), 241–265.

[317] Tao Yang and Qingyao Ai. 2021. Maximizing Marginal Fairness for Dynamic Learning to Rank. In *Proceedings of the Web Conference 2021*. 137–145.

[318] Zhihai Yang, Lin Xu, Zhongmin Cai, and Zongben Xu. 2016. Re-scale AdaBoost for attack detection in collaborative filtering recommender systems. *Knowledge-Based Systems* 100 (2016), 74–88.

[319] Sirui Yao and Bert Huang. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. In *Advances in Neural Information Processing Systems*.

[320] Hongtao Yu, Ruibo Gao, Kun Wang, and Fuzhi Zhang. 2017. A novel robust recommendation method based on kernel matrix factorization. *Journal of Intelligent & Fuzzy Systems* 32, 3 (2017), 2101–2109.

[321] Fajie Yuan, Alexandros Karatzoglou, Ioannis Arapakis, Joemon M Jose, and Xiangnan He. 2019. A simple convolutional generative network for next item recommendation. In *Proceedings of the Twelfth ACM International Conference on*

Web Search and Data Mining. 582–590.

[322] Feng Yuan, Lina Yao, and Boualem Benatallah. 2018. Adversarial Collaborative Auto-encoder for Top-N Recommendation. *arXiv preprint arXiv:1808.05361* (2018).

[323] Feng Yuan, Lina Yao, and Boualem Benatallah. 2019. Adversarial collaborative neural network for robust recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1065–1068.

[324] Huanhuan Yuan, Jian Yang, Jiajin Huang, and Ning Zhong. 2021. Probabilistic Embedding for Collaborative Reasoning. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 25–31.

[325] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW*. 1171–1180.

[326] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. 2019. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research* 20, 1 (2019), 2737–2778.

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

[328] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2021. Fairness in ranking: A survey. *arXiv preprint arXiv:2103.14000* (2021).

[329] Justin Zhan, Chia-Lung Hsieh, I-Cheng Wang, Tsan-Sheng Hsu, Churn-Jung Liau, and Da-Wei Wang. 2010. Privacy-preserving collaborative recommender systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 40, 4 (2010), 472–476.

[330] Fuzhi Zhang, Yuanli Lu, Jianmin Chen, Shaoshuai Liu, and Zhoujun Ling. 2017. Robust collaborative filtering based on non-negative matrix factorization and R1-norm. *Knowledge-based systems* 118 (2017), 177–190.

[331] Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A Survey of Controllable Text Generation using Transformer-based Pre-trained Language Models. *arXiv preprint arXiv:2201.05337* (2022).

[332] Junzhe Zhang and Elias Bareinboim. 2018. Equality of opportunity in classification: A causal approach. In *NeurIPS*. 3675–3685.

[333] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *AAAI*, Vol. 32.

[334] Minxing Zhang, Zhaochun Ren, Zihan Wang, Pengjie Ren, Zhunmin Chen, Pengfei Hu, and Yang Zhang. 2021. Membership Inference Attacks Against Recommender Systems. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 864–879.

[335] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* (2022).

[336] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)* 52, 1 (2019), 1–38.

[337] Wei Zhang, Junbing Yan, Zhuo Wang, and Jianyong Wang. 2022. Neuro-Symbolic Interpretable Collaborative Filtering for Attribute-based Recommendation. In *Proceedings of the ACM Web Conference 2022*. 3229–3238.

[338] Xin Zhang, Zengmao Wang, and Bo Du. 2022. Graph-aware collaborative reasoning for click-through rate prediction. *World Wide Web* (2022), 1–21.

[339] Yongfeng Zhang, Qingyao Ai, Xu Chen, and W Bruce Croft. 2017. Joint representation learning for top-n recommendation with heterogeneous information sources. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 1449–1458.

[340] Yongfeng Zhang and Xu Chen. 2020. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval* 14, 1 (2020), 1–101.

[341] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. 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*. 83–92.

[342] Yongfeng Zhang, Yunzhi Tan, Min Zhang, Yiqun Liu, Tat-Seng Chua, and Shaoping Ma. 2015. Catch the black sheep: unified framework for shilling attack detection based on fraudulent action propagation. In *Twenty-fourth international joint conference on artificial intelligence*.

[343] Yongfeng Zhang, Haochen Zhang, Min Zhang, Yiqun Liu, et al. 2014. Do Users Rate or Review? Boost Phrase-level Sentiment Labeling with Review-level Sentiment Classification. *SIGIR* (2014), 1027–1030.

[344] Zhuo Zhang and Sanjeev R Kulkarni. 2014. Detection of shilling attacks in recommender systems via spectral clustering. In *17th International Conference on Information Fusion (FUSION)*. IEEE, 1–8.

[345] Lei Zheng, Vahid Noroozi, and Philip S. Yu. 2017. Joint deep modeling of users and items using reviews for recommendation. In *WSDM*.

[346] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1893–1902.

[347] Zirui Zhou, Lingyang Chu, Changxin Liu, Lanjun Wang, Jian Pei, and Yong Zhang. 2021. Towards Fair Federated Learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 4100–4101.

[348] Tianqing Zhu, Yongli Ren, Wanlei Zhou, Jia Rong, and Ping Xiong. 2014. An effective privacy preserving algorithm for neighborhood-based collaborative filtering. *Future Generation Computer Systems* 36 (2014), 142–155.

[349] Yaxin Zhu, Yikun Xian, Zuohui Fu, Gerard de Melo, and Yongfeng Zhang. 2021. Faithfully explainable recommendation via neural logic reasoning. *arXiv preprint arXiv:2104.07869* (2021).

[350] Chiara Zucco, Huizhi Liang, Giuseppe Di Fatta, and Mario Cannataro. 2018. Explainable sentiment analysis with applications in medicine. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 1740–1747.