Trustworthy Recommender Systems

SHOUJIN WANG, University of Technology Sydney, RMIT University, Australia
XIUZHEN ZHANG, RMIT University, Australia
YAN WANG, Macquarie University, Australia
FRANCESCO RICCI, Free University of Bozen-Bolzano, Italy

Recommender systems (RSs) aim at helping users to effectively retrieve items of their interests from a large catalogue. For a quite long time, researchers and practitioners have been focusing on developing accurate RSs. Recent years have witnessed an increasing number of threats to RSs, coming from attacks, system and user generated noise, and various types of biases. As a result, it has become clear that the focus on RS accuracy is too narrow, and the research must consider other important factors, particularly trustworthiness. A trustworthy recommender system (TRS) should not only be accurate but also transparent, unbiased, fair, and robust to noise and attacks. These observations actually led to a paradigm shift of the research on RSs: from accuracy-oriented RSs to TRSs. However, there is a lack of a systematic overview and discussion of the literature in this novel and fast-developing field of TRSs. To this end, in this article, we provide an overview of TRSs, including a discussion of the motivation and basic concepts of TRSs, a presentation of the challenges in building TRSs, and a perspective on the future directions in this area. We also provide a novel conceptual framework to support the construction of TRSs.

CCS Concepts: • **Information systems** → **Data mining**; • **Computing methodologies** → **Machine learning;**

Additional Key Words and Phrases: Recommender systems, trustworthy recommendation, trustworthy AI

ACM Reference format:
Shoujin Wang, Xiuzhen Zhang, Yan Wang, and Francesco Ricci. 2024. Trustworthy Recommender Systems. *ACM Trans. Intell. Syst. Technol.* 15, 4, Article 84 (July 2024), 20 pages.
https://doi.org/10.1145/3627826

1 INTRODUCTION

We are living in the era of information explosion and digital economy, where the information overload problem has become increasingly important. As a matter of fact, today people often make choices from massive and rapidly growing catalogues of products and services (generally called *items*), while consuming a large amount of time and resources to discriminate relevant from

This work was supported by Australian Research Council Discovery Projects DP200101441 and DP230100676, the 2023 UTS Key Technology Partnerships Seed Funding Scheme, and the 2023 UTS MCR Research Capabilities Development Initiative.
Authors' addresses: S. Wang, FEIT, University of Technology Sydney, Sydney, NSW 2007, Australia; e-mail: shoujin.wang@uts.edu.au; X. Zhang (Corresponding author), School of Computing Technologies, RMIT University, 124 La Trobe Street, Melbourne, VIC 3000, Australia; e-mail: xiuzhen.zhang@rmit.edu.au; Y. Wang (Corresponding author), School of Computing, Macquarie University, Balaclava Rd, Macquarie Park NSW 2109, Australia; e-mail: yan.wang@mq.edu.au; F. Ricci, Free University of Bozen-Bolzano, Piazza Università, 1, 39100 Bozen-Bolzano, Italy; e-mail: fmr959@gmail.com.

2157-6904/2024/07-ART84 $15.00
https://doi.org/10.1145/3627826

not-relevant items. To make informed choices and decisions in a more effective and efficient way, **Recommender Systems (RSs)** have been introduced into almost every aspect of our daily life, work, business, study, entertainment and socialization [53, 64]. Today, RSs are one of the most important and popular application areas of **Artificial Intelligence (AI)**. RSs are software tools and techniques which provide suggestions on items which may be of interest to end users. According to McKinsey's report,[1] 35% of what customers purchase on Amazon and 75% of what users watch on Netflix come from recommendations.

Since the release of the first RS, "Tapestry" in 1992, which was then termed the *collaborative filtering technique* [22], RSs have flourished and achieved great success in the past 30 years. A series of RS approaches and models including content-based filtering, collaborative filtering and hybrid approaches have been developed and deployed, most of which have achieved good recommendation performance. In recent years, benefiting from the advancement of machine learning, especially deep learning, more powerful and accurate RS models have been proposed [53, 85].

Although tremendous successes have been achieved in the RS area, the majority of existing research works has focused on the improvement of the system accuracy, while only some minor part of the literature has also taken some other important properties and values of RSs into account, such as diversity and novelty [7]. In fact, existing accuracy-oriented RSs have not properly considered end users in the broader context of human-machine interaction where other important factors are commonly considered as critical for the overall user experience [35]. There are some additional aspects that require particular attention. Firstly, the cyberspace where RSs are deployed is becoming increasingly complex and subject to threats coming from various sources and of diverse nature, including cyber-attacks [11], noisy and fake information [70], and system bias [9]. This triggers the demand of **Trustworthy Recommender Systems (TRSs)** which aim to better serve users in the complex and challenging cyberspace. Secondly, stakeholders, including users, owners and regulators of RSs, have increasingly higher demand for RSs [1]. These stakeholders not only demand recommendation accuracy but also need trustworthiness, including robustness, fairness, explainability and privacy preservation. Actually, trustworthiness is even more important than accuracy in some critical and sensitive domains, including finance and medicine, where highly reliable RSs are required. In practice, it has become a consensus both in the academia and industry that accuracy should not be the only focus of an RS, and trustworthiness must be prioritised. These analyses have triggered the urgent demand of a new RS paradigm (i.e., TRSs).

Some researchers have already started to analyse the existing work in the area of TRSs. For instance, Ge et al. [19] have systematically surveyed the techniques for key aspects related to trustworthy recommendation, including explainability, fairness, privacy-preserving, robustness and user controllability in recommendation. Similarly, Fan et al. [18] have conducted a comprehensive overview of TRSs, specially focusing on six important aspects, including safety and robustness, nondiscrimination and fairness, explainability, privacy, environmental well-being, and accountability and auditability. Jha et al. [33] have presented an exhaustive survey to emphasise the current research in the field on TRS models. Jin et al. [34] have surveyed existing methodologies and practices of fairness in RSs, which is an important aspect of TRSs. Zhang et al. [87] have performed a comprehensive survey on explainable recommendations, by considering information sources for explanations, explanation models, explanation evaluations and applications. These surveys focus on summarizing existing work and emphasising the progress in one [34, 87] or multiple [18, 19, 33] aspects in the field of TRS. However, they lack a high-level overview of the subject and do not propose a unified framework for TRSs. These two aspects differentiate them from our work. Moreover, instead of summarizing progress in the literature, this work aims to focus on some new

[1] https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers

perspectives on the overall landscape of TRSs, featured with a unified framework to well connect and organize all different aspects of TRSs in a holistic way.

In practice, we believe that researchers are still lacking a systematic overview of TRSs: there is neither a work that systematically discusses the fundamental concepts or critical challenges nor a contribution that provides a unified framework for TRSs. Both aspects are quite important for the further development of TRSs. Therefore, it is urgent to provide a high-level overview on TRSs to comprehensively treat the fundamental concepts and key challenges and to give a broad picture of this emerging area.

The rest of the article is organized as follows. In Section 2, we will explore the fundamental concepts underlying TRSs, including the novel proposal of trustworthy recommendation ecosystem, aspects of TRSs and examples of TRSs. In Section 3, we will present a systematic review of the paradigm shift in the field of recommendation, followed by an illustration of a four-stage perspective for TRS in Section 4. In Section 5, we will systematically analyse the critical challenges faced in building TRSs in each of the four stages. In Section 6, we will present a conceptual framework to build TRSs and then will share some future directions in this vibrant area in Section 7. We will conclude this work in Section 8.

2 THE CONCEPT OF TRUSTWORTHY RECOMMENDER SYSTEMS

The term *trustworthy* describes an entity on which a subject can rely to be good, honest and sincere [14]. Specific aspects of "trustworthy" include trustable, reliable, dependable, faithful, honourable, creditworthy and responsible [83]. In general, a *TRS* refers to an RS on which its stakeholder can rely to be good, trustable, reliable, dependable and faithful.

A recommendation process is complex and interactive; it involves several entities such as users, items, the RS and even the provider of items (e.g., sellers). All of them actually form a *recommendation ecosystem* which enables the recommendation activities [1, 2]. In practice, these are the necessary elements for generating recommendations: data (i.e., user/item/provider-related information, e.g., user-item interactions), approaches and models (i.e., RS).

TRSs are actually the result of TRS approaches and models built in a trustworthy ecosystem. Although TRS approaches and models are pivotal for enabling trustworthy recommendations, they cannot survive without a trustworthy ecosystem. Most of the studies in the literature only discussed TRS approaches while ignoring the significance of trustworthy ecosystem. For instance, Ge et al. [19] provided a timely survey on the various techniques for enabling robust, fair, explainable or secure RSs. However, they have not examined trustworthy ecosystems or complex challenges in building TRSs. This greatly differentiates their analysis from our work. To bridge this gap, we first propose a novel concept of the trustworthy recommendation ecosystem, which is the foundation for building TRSs. In addition, a trustworthy recommendation ecosystem can well differentiate TRSs from the generally discussed trustworthy AI, since it is derived from the core characteristics and properties of the recommendation scenario. Then, we illustrate the various aspects to specify the "trustworthiness" of RSs with an emphasis on the trustworthiness of recommendation approaches and models. Finally, we will describe some specific and straightforward examples of TRSs in the real world.

2.1 Trustworthy Recommendation Ecosystem

A recommendation process unfolds in a complex user-RS-item interaction scenario. Items are provided by various providers, including sellers and news source websites, and thus providers are an indispensable part of the process [30]. As mentioned earlier, these four parties together form a basic ecosystem for recommendation activities, and they interact with each other, possibly in a collaborative form, to complete the recommendation task. A trustworthy recommendation

ecosystem includes trustworthy *users, items, providers* and *RSs*, which will be illustrated successively in the following paragraphs.

Trustworthy Users. Users consuming recommendations can possess diverse characteristics and perform various tasks in the cyberspace. Some users may not be trustworthy, and may try to obstacle the proper functioning of the RS. For instance, it is not uncommon to identify special users (e.g., internet water armies and internet hackers) who may attempt to perform fake or false interactions with items to intentionally bias the recommendation models for a given personal and even malicious purpose (e.g., biased product promotions) [86]. Therefore, to build a TRS, only trustworthy users should be authorised to operate. Trustworthy users can be roughly characterised by two aspects: (1) the user profile information should be true and accurate, and users' behaviour on the recommendation platforms should reflect genuine preferences and intentions [66, 67], and (2) users should maintain a good reputation in their interaction behaviour [72, 74] (e.g., they should write true and objective reviews or ratings for items, and pay for the purchased items on time).

Trustworthy Items. Usually, there are massive quantities and diverse types of items on an online consumption platform. Items have diverse characteristics and can bring different impact to the RS stakeholders, which include individual users, the local community and even the whole society. Although most of the items can well satisfy users' needs and can have a positive impact on the stakeholders' interests in the long run, some items may produce a negative impact on either the individual users, the community or the society. For example, additive electronic games usually match young children's preferences well. But these games often create addiction and thus can negatively affect users' lives, work or studies, hence resulting in significant negative impact on users and their family. Another example is fake news, which may match some minority users' stance and reading preferences. But they mislead the public and may cause serious social problems, and should not be recommended to end users [70]. To this end, only trustworthy items should be recommended by a TRS. Trustworthy items can be characterised by two aspects: (1) the item-related information (e.g., item attributes) used for recommendations should be true and precise so that the intrinsic characteristics of all items can be easily captured, and (2) the candidate items to be recommended to end users should be trustable and responsible, namely the items must not bring potential negative impacts to the stakeholders of RSs. Hence, true and unbiased news articles [63, 70], and healthy food, are such examples.

Trustworthy Providers. Providers are the party who provide items in online platforms. For example, a provider may be a seller in an e-commerce platform or a news source website which releases original news in the news domain. In the real world, some providers are not trustworthy since they may provide fake or poor-quality items, leading to negative impact to the end users and the society. For instance, some sellers may sell fake products, whereas some news websites may release fake news. Obviously, trustworthy providers are necessary for building TRSs, and they can be characterised by two aspects: (1) the providers should be responsible for the quality and potential impact of the items they provide. For instance, the sellers should provide products or services with guaranteed quality, whereas the news websites should provide verified true news only while combating fake news, and (2) it is important for providers to maintain a good reputation through their historical transaction or releasing behaviours [71, 82]. For example, those reputable sellers and providers who provide high-quality items in a responsible way without any harm to individuals and the society [68], and receive true and good reviews from their customers, are trustworthy providers.

Trustworthy Recommender Systems. Recommendation approaches, which are usually built by using data mining or machine learning models, should be reliable and stable. Specifically, they

should not only be able to accurately capture users' preferences and items' characteristics, so as to provide accurate recommendation services to end users, but also perform stably all the time, even in the most complex, dynamic and challenging scenarios (e.g., the cyberspace facing frequent cyber-attacks). Trustworthy approaches and models are the most important part in TRSs, and they can be characterised by multiple specific aspects, which will be discussed in detail in the next section.

2.2 Aspects of TRSs

In some related areas, such as trustworthy AI, many different aspects have been proposed to specify "trustworthy" from perspectives including regulations, mechanism and technology [83]. Several aspects related to the technology perspective have been commonly recognised, including robustness, fairness, explainability, transparency, privacy, accountability and responsibility. Since RSs are a specific application area of AI, the aspects used to define trustworthy AI are also applicable to characterise TRSs. However, due to the very special and specific data characteristics, and functioning and computation task, of an RS, there are additional aspects, such as a human's perception and trustworthy evaluation, which are specific to the RS scenario and distinguish TRSs from trustworthy AI. Next, we illustrate various features of TRSs:

- *Robustness* indicates an RS that has strong fault-tolerant capability to survive attacks, and noisy and fake information (e.g., users' arbitrary behaviours and fake reviews/ratings on items) [39, 48], which often commonly affect the input data used for recommendations.
- *Fairness* indicates that the RS is able to reduce or remove possible bias towards certain groups of stakeholders (e.g., the disparities in treating either user with different demographics or items with low and high popularity) which may be present in the input data, the recommendation model or in the recommendation results [20, 46, 73]. Consequently, the final recommendation results can be assessed as fair from both the user side and the provider side.
- *Transparency* generally means that the functioning mechanism and the recommendation model are transparent to all the stakeholders of the RS, and it is not perceived as a "black box" [6, 13]. Transparency can greatly reduce the potential risks of deploying ineffective RSs in real-world domains, especially in some sensitive and critical domains like finance and healthcare domains.
- *Explainability* means that the recommendation mechanism, recommendation models and the recommendation results are well explained so that the stakeholders of the RS can properly understand how and why the recommendations are generated [38, 61, 87].
- *Privacy and security* means that the RS is able to effectively protect the personal privacy of relevant stakeholders [28, 58]. This aspect is even more important in RSs, compared to other web applications, since RSs are consuming a large amount of profile and behaviour, which often contains a sensible users' characteristics.
- *Responsibility*: The development and deployment of RSs should be conducted in a responsible manner (e.g., follow the regulations and law) so that the RS will not explicitly or implicitly harm anybody. Moreover, the generated recommendation results should be beneficial to all stakeholders [17, 32]. For instance, fake news should not be recommended to any users even if it may interest some, or addictive games should not be recommended to young children either.
- *Human's perception of trustworthiness* relates to the subjective perception, which the relevant stakeholders have, of the trustworthiness of the RS. This is unique to RSs and is of great significance. In fact, an RS is actually involving a human-machine interaction [31], and whether an RS is trustworthy or not ultimately depends on a human's feeling and perception. However, almost all of the existing approaches to model TRSs have been developed by assuming a single point of view, that of the machine [77] (i.e., the RS models and algorithms), but have

ignored the human factor. Moreover, existing work often mechanically computes a numerical measurement value to quantify the trustworthiness of RSs. Such lack of consideration of the human factor triggers the urgent demand to take human perception into account when talking about TRSs. For instance, in the real-world cases, if an RS frequently recommends those items similar or identical to those items which have been purchased by users very recently, then users may feel that the RS is not so trustworthy. This is because the RS results in a lot of duplicate recommendations in a short period, which do not match users' demand and preferences. Instead, if the RS can recommend some items different from but relevant to users' purchase history [59], then users may feel that it is relatively more trustworthy since it can better accommodate users' actual demand and preferences.

- *Trustworthiness integration*: Although various aspects and perspectives have been proposed to characterise TRSs, they are mostly separated from each other. In addition, most of the existing studies on TRSs have focused on one or two aspects of trustworthiness alone. For instance, fairness-aware RSs and explainable RSs only focus on the fairness and explainability aspect, respectively. However, accounting for only one or two aspects cannot lead to a truly trustworthy RS. To build truly trustworthy RSs, all the aforementioned different aspects, including both objective and subjective aspects, should be integrated together organically and effectively towards a unified TRS framework.
- *Trustworthy evaluations*: On one hand, both the evaluation methods and the evaluation metrics must be reliable—that is, they must well and precisely indicate the performance level of the RS. For instance, offline evaluations are often performed under very ideal situations, and thus they cannot indicate how the tested RS will perform in the real online scenario [23]. On the other hand, new evaluation protocols and metrics are in demand to evaluate the trustworthiness of an RS so that all the aforementioned aspects can be appropriately validated.

Although these different features of TRSs touch a variety of issues and often correspond to different elements and stages of an RS, they are closely inter-connected and must be jointly addressed to contribute to the "trustworthiness" of the RS, as illustrated in Figure 1. For instance, fairness often relates to explainability. In fact, the fairness perspective (e.g., fair treatment of users with different gender), when managed by the RS, should also be explainable to the end users so that they can clearly understand and accept the generated fair recommendation results.

2.3 Examples of TRSs

So far, RSs have been widely exploited in nearly all information systems that we use daily for work, study or entertainment. In recent years, along with the occurrence of an increasing number of cyber threats including cyber-attacks, fake or false online information, and bias in the cyberspace, TRSs are an evident need. There is a variety of examples of real-world TRSs across different application domains and sectors.

In the e-commerce domain and other product-based sectors, TRSs are of great significance for recommending products in a responsible way that users can trust. One example is the robust RS which can tame fake ratings and reviews [42] on platforms like eBay or Amazon. This type of RS detects fake ratings and reviews inserted on the platform and discards them when generating trustworthy recommendations.

In the media domain and other content-oriented sectors, such as news websites, social media and video websites, TRSs are essentially important for promoting content towards social good. A typical example is a recently proposed fact check technology (e.g., Google Fact Check tools) for news recommendations, which introduces a series of techniques to check the veracity of news on the web so that only verified true news will be recommended to end users [21]. Another typical example is a veracity-aware news RS published at the 2022 Web Conference [70]. This system

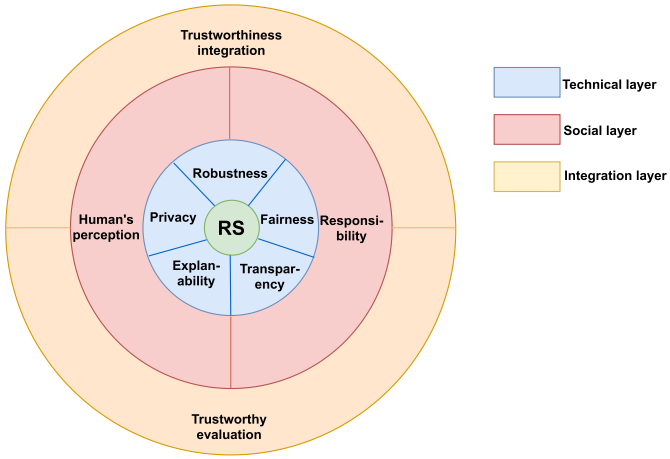

Fig. 1. The various aspects to characterise TRSs and the relations among them. According to the characterisation perspective, they can be organized into three layers.

checks the veracity of each news piece before generating recommendations so that only checked true news will be recommended to end users. Furthermore, on video platforms like YouTube [10], only those videos which are positive and can benefit end users and the society should be recommended. In contrast, those additive videos and games should not be recommended, especially to teenagers.

In the field of tourism and other service sectors, TRSs are critical to combat various bias from different sides, so as to provide fair and sustainable recommendations. For example, in point-of-interest recommendations, both user bias from various active levels of users and item popularity bias across short-head, mid-tail, and long-tail groups of items should be mitigated [45, 52]. Moreover, the sustainability of the generated recommendations and the impact of items' consumption produced by recommendations must be considered [47].

3 THE RECOMMENDATION PARADIGM SHIFT

3.1 From Accuracy-Oriented RSs to TRSs

RS research dates back to the 1990s, when the first collaborative system "Tapestry" was described in the literature [22]. From the early 1990s to the middle of the 2000s, several algorithms were developed for enabling content-based or collaborative filtering recommendations [53, 76]. From 2008 to 2016, driven by the popular Netflix challenge, matrix factorization methods dominated the scenario for a long time. Since the middle of the 2010s, benefiting from the rapid development of deep learning techniques, deep learning based RSs have become the mainstream in the recommendation domain [84, 85].

Along with that technological shift and revolution of RSs, the focused research problems, challenges and evaluation mechanisms of RSs have also changed. Since the first studies on RSs to the early 2010s, the accuracy of recommendation results has been the most important and often the only evaluation criteria that was used to evaluate an RS. Specifically, that means that the researchers tried to assess whether the recommended items were truly preferred (e.g., clicked, purchased, highly rated) by users. During this period, nearly all efforts were devoted to develop more accurate RS models and algorithms, such as more accurate matrix factorization algorithms to predict the unknown user-item ratings with a smaller and smaller prediction error. Then, starting

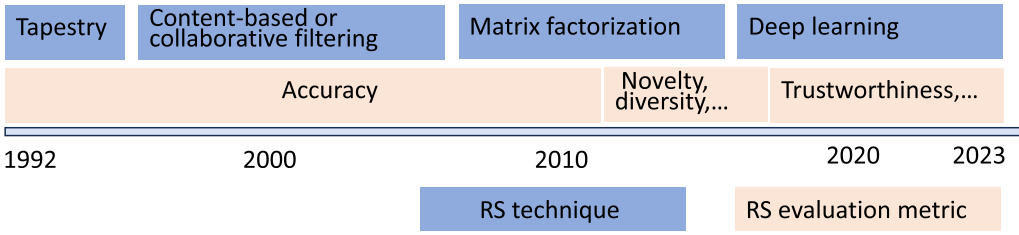

Fig. 2. The evolution paths of RS technique and RS evaluation.

from the early 2010s, more and more researchers realized that the accuracy criterion is insufficient for evaluating an RS, and thus new evaluation dimensions, including diversity and novelty, were proposed [4, 23]. As a result, the evaluation of an RS moved from single-criteria evaluations, based on accuracy, to multi-criteria evaluations, based on a range of dimensions, such as accuracy, diversity and novelty, whereas accuracy still was considered as the most important criteria.

Since the late 2010s, due to an increasing level of threats in the cyberspace, coming from cyber-attacks, noises and biases, the trustworthiness of RSs have attracted increasing attention from both academia and industry. As a result, a growing number of studies have focused on tTRSs from various perspectives such as robustness, fairness, explainability or privacy. For example, some work aimed to build robust RSs which can fight against shilling attacks [24] and fairness-aware RSs to generate fair recommendations for users with different demographics [78]. Some other work has instead focused on building explainable RSs to generate user-understandable recommendation results [87], or secure and privacy-preserved RSs which puts a special concern on users' privacy [58]. Obviously, the research focus and the ultimate goal of these studies is no longer centered on the improvement of recommendation accuracy. Instead, they have changed the direction towards enhancing the trustworthiness of RSs. This has actually revolutionised the research focus and the evaluation perspectives, from accuracy-oriented recommendations towards trustworthiness-oriented recommendations, leading to a recommendation paradigm shift.

In practice, the trustworthiness of RSs has been attracting much attention from different sectors including academia, industry and government. For instance, the Chinese government released regulations on recommendation algorithms for internet information services in early 2022 [50]. Moreover, the Digital Service Act of the European Union has been published in the Official Journal since 27 October 2022 and came into force on 16 November 2022. It explicitly mention RSs and requires platform owners to adhere to regulations related to auditing, transparency and freedom of users to opt out from RSs. A large proportion of the regulations emphasised the significance of TRSs to be deployed in real-world business applications. The evolution paths along with time of both RS techniques and RS evaluations are demonstrated in Figure 2.

3.2 From Trust-Aware RSs to TRSs

Trust-aware RSs mainly consider and leverage the trust relationships between humans and organizations to enhance the recommendation performance. Here the trust relationships can vary, and some typical examples include the trust relations between different users on online or offline social networks (e.g., friends on Facebook, or colleagues in one organization), the trust relationships between users and comments, ratings on items from others, and the trust relationships between customers and sellers [81, 88]. Since the early 2000s, a variety of studies [5, 16, 44, 82] have investigated how to model and leverage such diverse trust relationships to improve the performance of RSs.

Differently from trust-aware RSs, TRSs aim to build reliable RSs which can be accepted by humans, which were proposed in the late 2000s for the first time [48]. However, early definitions on TRSs mainly focused on one single aspect of trustworthiness, namely robustness [48, 77]. So TRSs in the early stage mainly refer to RSs which are robust while facing various types of attacks in the cyberspace. However, there are some other important aspects contributing to trustworthiness that were initially overlooked, including the aforementioned fairness, explainability, privacy and responsibility.

Therefore, it is of great theoretical and practical significance to develop the next-generation TRSs which go beyond the existing robustness-oriented ones. To this end, in this article, we propose novel TRSs in which all the nine aforementioned aspects contributing to TRSs are considered, leading to a significant step towards the implementation of truly trustworthy RSs.

4 TRUSTWORTHY RECOMMENDER SYSTEMS: A FOUR-STAGE PERSPECTIVE

Classical Recommendation Process. Generally speaking, a typical machine learning based recommendation process comprises the following four successive and inter-connected stages:

- *Data preparation stage*, including data collection and preprocessing;
- *Data representation stage*, which often includes learning an informative representation of the raw input data as the input of the downstream recommendation model;
- *Recommendation generation stage*, which requires some data mining or machine learning based prediction models to predict the unknown user-item interactions (e.g., ratings, clicks); and
- *Performance evaluation stage*, which employs various evaluation methods, including online and offline evaluations, to evaluate the recommendation results from different perspectives such as accuracy, diversity and fairness.

Trustworthy Recommender Systems. To be trustworthy, an RS should be trustworthy in each of the aforementioned four stages. Each stage has its unique computation task, characters and goal, so the "trustworthiness" of different stages often has different specific meanings and focuses on different aspects of the aforementioned nine aspects (cf. Section 2.2). For instance, the data preparation stage often focuses more on the "robustness" aspect, since the data preparation methods should be robust enough to survive to noisy and fake data, possible coming from attacks, whereas the recommendation generation stage should focus more on aspects like "fairness", "transparency" and "robustness", since fair, transparent and robust RS models and algorithms are required in this stage for generating recommendations. Next, we try to characterise a TRS from a four-stage perspective while the specific meaning of "trustworthy" in each stage will be highlighted:

- *Robust and secure data preparation*, on one hand, is necessary to combat noisy and fake information, and bias in the raw data, generated in the cyberspace, and survive possible attacks on the data. In this way, the prepared data for recommendations can accurately reflect users' preferences and items' characteristics. On the other hand, the users' data should be managed and processed in a secure way so that sensitive and privacy information will not be leaked.
- *Robust and explainable data representation* employs robust and explainable data representation models to precisely learn informative and interpretable latent representations of the original input data in a stable way regardless of the potential attacks. In most cases, the learned latent representations are latent vectors without an explicit semantic meaning, which then negatively impacts on explainability. In addition, it is often not very clear which information is encoded in the representations. Therefore, the "explainability" aspect should be emphasised in this stage.

- In *fair, transparent, explainable and responsible recommendation generation*, fair, transparent and explainable RS models should be adopted to provide fair exposure opportunities of all items to different users in a transparent, explainable and responsible way. In addition, the models should also be able to provide straightforward explanations by using texts or images corresponding to the recommendation results, to faithfully explain how the recommendations are generated and why they are appropriate for a given user. More importantly, only those items with positive impact to stakeholders can be recommended.
- *Trustworthy evaluations* comprise two-sided evaluations: (1) *technical evaluation* and (2) *ethical evaluation*. The technical evaluation aims to evaluate the recommendation performance from some technical perspectives such as accuracy, diversity, novelty and explainability. More trustworthy and reliable evaluation protocols and metrics are required to evaluate such technical performance. In contrast, the ethical evaluation aims to evaluate recommendations from the ethical perspective, such as the responsibility and social impact—specifically, whether the recommendation activities and the recommendation results are responsible and beneficial to the relevant stakeholders. For instance, the recommendations of fake news and misinformation on the web is not ethical and may even be harmful to society due to their possible vetted interest (e.g., political propaganda). Towards TRSs, ethically concerned evaluations are necessary and of great significance to the sustainable development of the community and the society. Some examples of unethical but profitable recommendations include the recommendations of fake news [70] and the recommendation of addictive games to young children. To the best of our knowledge, this is the first time the proposal of an ethical evaluation of RSs is put forward in the literature.

5 CHALLENGES FOR BUILDING TRUSTWORTHY RECOMMENDER SYSTEMS

Although the significance and urgency of TRSs have been recognised, the research in this area is still in an early stage. A variety of challenges and research problems remain to be properly addressed in each of the aforementioned four stages, as illustrated in Figure 3. We therefore analyse the particular challenges of each stage.

Challenges in the Data Preparation Stage. Before building an RS model, a necessary step is to collect and prepare data to train and test the model [69]. Such data often contains user-item interactions such as users' clicks, purchases, ratings or comments on items, and side information such as users' attributes, social relations and items' features [29, 65]. The data is often collected from online consumption platforms such as amazon.com, youtube.com, and abcnews.go.com. Due to the large amount of diverse users (e.g., 10M+ users) on each platform, it is inevitable that some users' behaviours may be noisy and unreliable (e.g., randomly click or view a set of items on amazon.com), or there may be some malicious users (e.g., internet water armies) who have performed malicious behaviours, such as posting fake ratings or comments to some targeted items or sellers. These factors bring noise, bias and fake information into the interactions or side information, making the data used for recommendations unclean and unreliable. As a result, the data cannot accurately reveal the users' preferences and item characteristics, and thus misleads the downstream recommendation models.

Therefore, a big challenge at the data preparation stage is to effectively detect and remove unreliable data, such as noisy, fake or biased data, and retain the genuine data. This task is quite challenging since this critical data usually accounts for a quite small proportion of the whole population, making it difficult to discover it, especially when unreliable data looks quite similar to the genuine data. Possible strategies to address this challenge include a set of conventional and emerging data preprocessing techniques. For instance, some advanced data augmentation and data debias

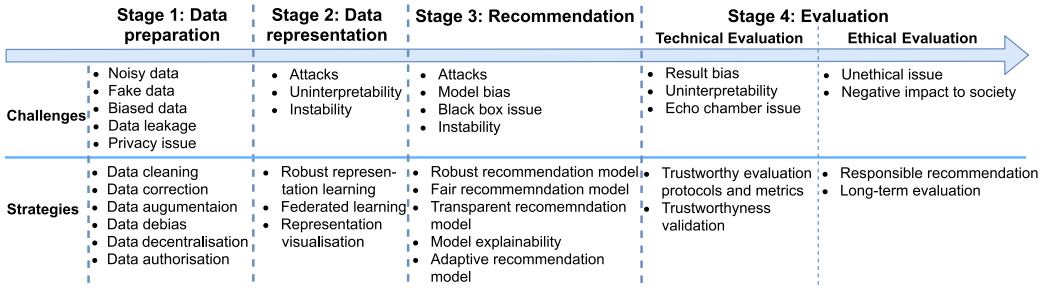

Fig. 3. The challenges and corresponding strategies for building TRSs.

methods [56], including counterfactual data augmentation [75] and contrastive learning [60], have been extensively studied to remove the bias in raw data in recent years.

Another challenge at this stage is related to the data privacy and security issue. This is especially important in real-world RS applications since the data used for recommendations often contains a large amount of users' private information which should be protected. To be specific, on one hand, during the collection of raw data, it is important to ensure that only the necessary data will be collected. On the other hand, when a large amount of data is collected, it is important to build a safe and strong infrastructure and mechanism to properly store and manage it, including a secure and powerful database, and a rigorous data access mechanism. Another way to address the data privacy and security issue is to introduce federated learning [80] and edge computing [57]. In this way, the data used for recommendations can be decentralized to various edge devices such as users' mobile phones where the individual users' privacy can be well protected.

Challenges in the Data Representation Learning Stage. Data representation aims to learn an informative data model, such as a numerical vector in a latent space, to well represent each data point in the original input dataset. For instance, the well-known representation learning model Word2Vec [54] learns a latent numerical low-dimensional vector to represent each word.

In advanced machine learning, especially in the deep learning area, data representation learning is a prior step of all the downstream learning models and tasks. The quality of data representation directly determines the performance of the learning models. Therefore, machine learning based RS models often consist of two main steps: data representation learning and prediction. The output of the first step is taken as input of the second for the prediction task. For instance, in matrix factorization or neural network based RS models, each user/item is represented by a K-dimensional latent vector which is learned by mapping the user/item ID into a latent space [49]. The main challenges in the representation learning stage lie in the following three aspects.

First, how do we learn accurate and informative representations to precisely retain the original information in the raw data? For instance, the similarities/distances between the raw data points should be accurately reflected by their corresponding representations in the latent space. Towards this challenge, some advanced representation learning techniques, including contrastive representation learning [79], are able to learn more accurate and informative representations.

Second, how do we learn interpretable representations? One common and critical issue with latent representations is the lack of explainability. It is often not clear what the semantic meaning of the latent vectors is, making the learned representations uninterpretable for humans. Sometimes, it is not clear what kind of information is encoded in the latent vectors either, which brings some potential risks for the real-world applications of RS models. To address this challenge, some explainable representation learning models and techniques, including data visualization, can be developed to disclose the patterns encoded in the latent representations.

Third, how do we combat attacks on representation learning models? In general, nearly all machine learning models, including representation learning models, may suffer from some external attacks, especially those deployed in real-world applications. The attackers may quickly generate a large amount of fake data to intentionally mislead the models and thus downgrade the quality of the learned representations. Some possible strategies include robust representation learning techniques, such as adversarial representation learning which fights against attacks in an adversarial way [15].

Challenges in the Recommendation Generation Stage. In this stage, the data representations learned in the previous stage will be imported into the selected RS model for generating recommendations. The key challenge lies in how to build a *robust, fair, transparent, explainable* and *responsible* RS model which can work in a trustworthy way in a complex and dynamic environment.

First, similarly to what was observed for the representation learning stage, RS models also easily suffer from various attacks. So the first challenge is how to build robust RS models which can survive various attacks, especially those malicious attacks and adversarial attacks in the complex cyberspace where RSs are deployed.

Second, some particular design and work mechanisms of the RS models may lead to biased RS models and thus generate biased recommendations. Therefore, another challenge in this stage lies in how to effectively remove the bias. Various model debias techniques like auto-debias which learns the debias parameter from the data [8] can address this issue to some degree.

Third, advanced machine learning based RS models, especially deep learning based ones, often run in a black box mode, and thus the computations in the model are not transparent and explainable to humans. This makes it hard to understand how an RS model is working and what potential risks and disadvantages the model may have. Hence, this may bring potential risks for the applications of RS models in the real world, especially in some social and economic domains which are critical, such as healthcare and finance, where a small mistake may lead to a large loss (e.g., the loss of a life or a large sum of money). Hence, the third challenge in this stage lies in how to build transparent and explainable RS models whose work mechanisms are well understandable to humans. In this way, how and why the recommendations are generated can be well disclosed and visible to end users, which greatly reduces the risks of the employment of RSs in real-world business.

Fourth, considering the real-world application context, some recommendation behaviours may be profit-driven while ignoring the ethical or even the legal issues (e.g., recommending additive games to children). Some recommendation results may be accurate but not responsible, such as recommendations of fake news to particular users [70]. Hence, how to ensure responsible recommendation behaviours and recommendation results which are for social good is another big challenge in this stage.

Challenges in the Evaluation Stage. In this stage, the recommendation results, such as predicted user-item rating values or a list of ranked items, generated by an RS model are evaluated to measure its performance. Conventionally, the evaluation metrics could be a variety of quantity metrics of concern, such as accuracy, and diversity, which can directly measure the performance from the technical perspective. However, as mentioned in Section 1, trustworthy evaluation should go beyond conventional technical performance evaluation. Instead, they should contain both *technical evaluation* and *ethical evaluation*.

For the technical evaluation, existing evaluation protocols and metrics for RSs are mostly accuracy oriented [76]. They are not sufficient to comprehensively evaluate an RS, especially in the complex cyberspace which often contains noises, attacks and bias [27]. In such a complex context, the stakeholders of an RS usually focus not only on accuracy but also on trustworthiness. Hence,

the effective evaluation of the trustworthiness of an RS has great social and economic significance. However, the evaluation of the trustworthiness of an RS is quite challenging and involves multiple aspects such as robustness, fairness, explainability and privacy. Some of them are hard to be quantified in a traditional online or offline experiment. In addition, different application scenarios often have different requirements on the trustworthiness of the RS and focus on different aspects. For instance, in an online e-commerce platform where attacks and noise behaviours are routinely detected, the robustness of the RS should be a primary attention point. Conversely, in an offline RS for medicine treatment, the transparency of the RS model and the explainability of the recommendation results may be more important. Therefore, it is of great importance to define novel trustworthiness evaluation schemes for RSs, including both new evaluation protocols and evaluation metrics for comprehensively and systematically measuring the trustworthiness of RSs.

Ethical evaluation aims to evaluate an RS from ethical and social perspectives. Specifically, they measure whether the recommendation behaviours and recommendation results will have a positive impact on the various stakeholders. Here, the impact of an RS could be on the recognition or the behaviours of end users, and it could be explicit or implicit, short-term or long-term. Ethical evaluations are quite challenging since the various impact and influence of RSs on both individuals and society are very difficult to accurately capture and quantify [25]. New and well-designed ethical evaluation methods are in demand for RSs, which may combine both qualitative and quantitative evaluations.

6 A FRAMEWORK FOR TRUSTWORTHY RECOMMENDER SYSTEMS

To build a TRS is therefore a challenging task. Not only various aspects of trustworthiness, such as robustness, fairness and security, and their complex relations, should be considered in a unified way, but also new and trustworthy evaluation protocols and metrics are in order. Each aspect often involves specific models and techniques. For instance, robustness is often achieved via adversarial learning techniques such as generative adversarial networks [12]. Due to its extreme complexity, it is challenging to design a very specific model for TRSs which can handle all the aspects. However, a systematic methodology to build a comprehensive TRS while considering all the different aspects in different stages is a compelling challenge, and there is a need for a unified framework for TRSs. Hence, to bridge this gap, we propose a conceptual framework to support TRSs.

Consistent with the four stages of recommendation process discussed in Section 4, our proposed TRS framework consists of four stages: (1) *trustworthy data preparation*, (2) *robust and explainable data representation*, (3) *fair, transparent, explainable and responsible recommendation generation*, and (4) *trustworthy evaluation*, as depicted in Figure 4.

Trustworthy Data Preparation. Trustworthy data preparation aims to first collect raw data (e.g., user-item interaction data and side information data) from online platforms in a trustworthy way (e.g., privacy-preserved) and then transfer it to trustworthy input data (e.g., clean, unbiased and reliable data) for the downstream recommendation tasks. Note that the collected raw data often contains noises, bias and false information, among others. Some conventional and advanced data preprocessing techniques can be employed to support trustworthy data preparation. For instance, some data cleaning methods including pattern-based error detection [26] and machine learning based outlier detection [51] can be utilized to detect and remove those noisy and fake data in the raw dataset. In recent years, some advanced machine learning approaches like counterfactual data augmentation [75] and contrastive learning based data augmentation [60] have been commonly utilized to reduce the bias in input data for recommendations.

Robust and Explainable Data Representation. In principle, there are two specific steps for achieving trustworthy data representations: (1) representation learning to learn user and item

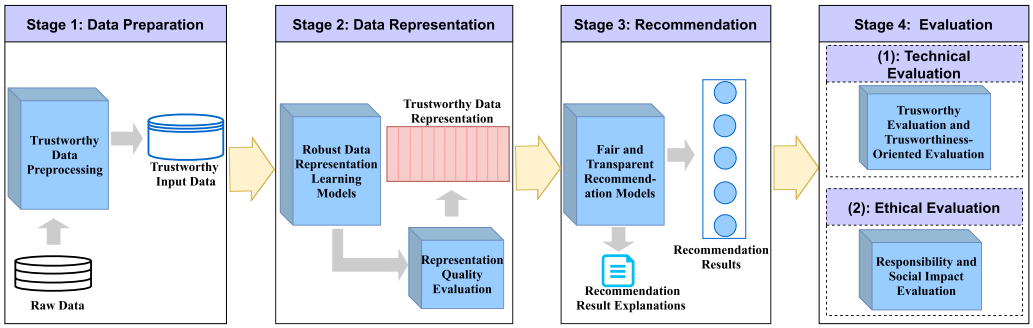

Fig. 4. A conceptual framework for building TRSs.

representations, and (2) representation quality evaluation to evaluate the learned representations. In the RS area, the evaluation step is often incorporated into the downstream recommendation task—that is, the representation quality is measured by the quality of recommendation results.

For the representation learning step, as discussed in Section 4, "trustworthy" can be further specified into robust and explainable. On one hand, to support the robustness of data representation learning models so that they can perform stably when facing attacks and noises, one typical solution is adversarial representation learning models [15]. In addition, to remove the possible biased and noisy information hidden in the representations, some representative debiasing and denoising models including Denoising Autoencoder (DAE) [41] can be deployed for more accurate representation learning for recommendations. On the other hand, to make the learned latent representations more interpretable for humans, some data visualization techniques such as t-SNE [62] can be helpful to explicitly reveal the hidden patterns in latent data.

Fair, Transparent, Explainable and Responsible Recommendation Generation. In the recommendation generation stage, the core task is to design and develop TRS models. As discussed in Section 4, "trustworthy" in this stage can be characterised along four dimensions: fair, transparent, explainable and responsible. We illustrate them in the following.

First, a TRS model must be free of bias so that it can perform equally well for different users and items. To support this, disentanglement learning based approaches [43] can be used for removing bias in RS models. In addition to RS models, bias may also exist in recommendation results, which could be reduced by re-ranking techniques based on a specific target constraint.

Second, TRS models should be transparent. To support the transparency, some set-based techniques can be employed to explicitly present the RS models to end users via natural language [6, 40]. Building transparent RS models is quite challenging, and more studies must address this topic.

Third, to support the explainability of recommendation results, the output of the recommendation generation stage should be two sided: (1) the normal recommendation results, such as a list of selected candidate items, and (2) corresponding explanations of the recommendation results to well explain how and why the recommendations are generated. The explanations should be straightforward and well understandable, either in textual or visual formats.

Last but not the least, to be trustworthy, the recommendation generation process must be responsible in terms of both the recommendation behaviours and the recommendation results. On one hand, some additional constraints and criteria can be employed to ensure the preceding aspects. For instance, some models can be developed to filter out fake products, additive games and fake news from recommendation lists. On the other hand, some regulations may be enforced to guide the harmonic development and deployment of RS techniques.

Trustworthy Evaluation. Evaluation is an important step for ensuring the quality of recommendation results generated by RS models. As depicted in Figure 4 and discussed in Section 4, both technical evaluation and ethical evaluation are required for a TRS. Regarding technical evaluation, most of the existing evaluations are accuracy oriented, whereas trustworthiness evaluations are in their early stage. Although a few metrics have been defined to measure one specific aspect of trustworthiness, such as fairness [37], there is no unified method to systematically evaluate multiple aspects as discussed in Section 2. New evaluation protocols and metrics are in demand for comprehensive trustworthiness evaluations.

Regarding ethical evaluation, to the best of our knowledge, there is no existing work reported in the literature. Ethical evaluation mainly evaluates the social impact of recommendation behaviours and results—for instance, whether the recommendation behaviours are legal and ethical or if the recommendation results lead to some good/bad impact (e.g., lead users to become addicted to video games). Ethical evaluation is full of complexity and dynamics, in which the users' cognition, perception and behaviours should be taken into account. To this end, some possible solutions include user studies, and the combination of qualitative and quantitative studies.

7 FUTURE DIRECTIONS

Truly trustworthy RSs. Although there have been works aiming at building TRSs, nearly all of them focus on only one single aspect (e.g., fairness or explainability). For instance, some early work on TRSs only focuses on the robustness of the RS [48], aiming to build resilient RSs to combat attacks. More recently, some other works have focused on either explainability [87], security [36] or fairness [37]. However, to be truly trustworthy, usually all these various aspects (cf. Section 2) of trustworthiness should be carefully considered in a unified way. Therefore, more analyses are in demand to build truly trustworthy RSs.

Human-Centered TRSs. Although some studies have attempted to build TRSs, almost all of them only mechanically developed the so-called trustworthy models from the machine perspective while ignoring the human factor. However, whether an RS is trustworthy or not should be ultimately judged by the relevant stakeholders rather than by a formal metric. Therefore, a human's perception and judgement should be well incorporated when developing and evaluating a TRS, which deserves a deeper investigation.

Trustworthy Recommendation Ecosystems. As discussed in Section 2, TRSs cannot survive without a trustworthy recommendation ecosystem. However, almost all the existing work on trustworthy RSs only focus on the development of TRS models while ignoring other important elements, such as trustworthy items and providers, in the ecosystem. Therefore, more studies in the development of trustworthy recommendation ecosystems are in order.

Multi-Granular Fairness-Aware TRSs. Fairness is essentially multi-granular and exists at different levels [46], such as the high-level fairness between different groups and low-level fairness between different individuals within each group. A really fair RS should be fair in all the different granularities rather than only a single granularity, as done in most of the existing work. Hence, to comprehensively model multi-granular fairness is of great significance for building truly fair RSs.

Fine-Grained Personalized RSs. Although it is a common sense that personalization is the core aspect and the basis of RSs, the recommendation results of most existing RSs are not personalized enough, since they fit the majority users only, while unsatisfying minorities of niche users. Therefore, it is in high demand for designing innovative RS models and evaluation approaches to ensure fine-grained personalization of recommendation results which can satisfy nearly each user.

Responsible RSs. Most existing RSs are technique driven, are built on advanced machine learning techniques, and are evaluated from the technical perspective only. The social impact of RSs are hardly considered during the design, development and evaluation of RSs. For example, the

recommendation of true and unbiased news and information items would bring positive impact to the society, whereas that of fake or biased information items would be harmful to the society and thus leads to negative impact. Another example is that of the potential negative effect of bad behaviours induced by RSs in the tourism domain, where large masses of tourists are influenced to visit certain, already popular, points of interest [47]. In fact, RSs are not only a technical issue but also a social one. More studies are needed to explore how to well incorporate social impact and social science into the whole lifecycle of RSs.

Large Language Models for Trustworthy Recommendations. In recent years, large language models have been flourishing, and have shown great potential in a variety of tasks including recommendations. Naturally, researchers have been investigating how to generate recommendations using large language models [55, 69]. Although promising, the potential risks when utilizing them for recommendations cannot be overlooked. For example, in models pre-trained on a large amount of textual data on the web, it is unavoidable that some bias may be introduced into them from the training data [3]. Some other underlying risks include privacy and security issues, transparency issue and explainability issue. For example, an individual's data may be leaked when using pre-trained large models for recommendations. Therefore, this poses an urgent yet challenging research problem: how to utilize large language models to achieve trustworthy recommendations.

New Evaluation Protocols and Metrics. As discussed in Section 2, building TRSs triggers new challenges for RS evaluations. On one hand, new technical evaluation protocols and measures are in demand to comprehensively measure the trustworthiness of RSs. On the other hand, new ethical evaluation methods are needed to effectively measure the social impact of RSs.

8 CONCLUSION

The TRS is a challenging yet demanding topic, which is of both great theoretical and practical value. In this article, we have provided a comprehensive overview of this topic. We have characterised TRSs by thoroughly analysing various aspects which can be used to characterise the trustworthiness of RSs. We have provided a novel four-stage framework to systematically illustrate TRSs and analyse the corresponding challenges in each of the four stages in building TRSs. We have also described a conceptual framework to support TRSs and pointed out some future directions in this novel and important area. The research on TRSs is flourishing, and it is our hope that this work can provide readers with a comprehensive understanding of the key aspects, main challenges and key techniques when building TRSs, and shed some light on future studies.

ACKNOWLEDGMENTS

Professor Huan Liu from Arizona State University has provided some feedback on the earlier version of this work.

REFERENCES

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Janach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction* 30, 1 (2020), 127–158.

[2] Himan Abdollahpouri and Robin Burke. 2022. Multistakeholder recommender systems. In *Recommender Systems Handbook*. Springer US, 647–677.

[3] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-Muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 298–306.

[4] Gediminas Adomavicius and YoungOk Kwon. 2011. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering* 24, 5 (2011), 896–911.

[5] Milad Ahmadian, Mahmood Ahmadi, and Sajad Ahmadian. 2022. A reliable deep representation learning to improve trust-aware recommendation systems. *Expert Systems with Applications* 197 (2022), 116697.

[6] 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 (SIGIR'19)*. 265–274.
[7] Pablo Castells, Neil Hurley, and Saúl Vargas. 2022. Novelty and diversity in recommender systems. In *Recommender Systems Handbook*. Springer US, 603–646.
[8] Jiawei Chen, Hande Dong, Yang Qiu, Xiangnan He, Xin Xin, Liang Chen, Guli Lin, and Keping Yang. 2021. AutoDebias: Learning to debias for recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'21)*. 21–30.
[9] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.
[10] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, and Taylor Van Vleet. 2010. The YouTube video recommendation system. In *Proceedings of the 4th ACM Conference on Recommender Systems*. 293–296.
[11] Ignacio Fernandez De Arroyabe, Carlos F. A. Arranz, Marta F. Arroyabe, and Juan Carlos Fernandez de Arroyabe. 2023. Cybersecurity capabilities and cyber-attacks as drivers of investment in cybersecurity systems: A UK survey for 2018 and 2019. *Computers & Security* 124 (2023), 102954.
[12] 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* 54, 2 (2021), 1–38.
[13] Tommaso Di Noia, Nava Tintarev, Panagiota Fatourou, and Markus Schedl. 2022. Recommender systems under European AI regulations. *Communications of the ACM* 65, 4 (2022), 69–73.
[14] Oxford Learner's Dictionaries. 2022. Definition of Trustworthy. Retrieved October 20, 2023 from https://www.oxfordlearnersdictionaries.com/definition/american_english/trustworthy
[15] Jeff Donahue and Karen Simonyan. 2019. Large scale adversarial representation learning. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS'19)*. 10542–10552.
[16] Manqing Dong, Feng Yuan, Lina Yao, Xianzhi Wang, Xiwei Xu, and Liming Zhu. 2022. A survey for trust-aware recommender systems: A deep learning perspective. *Knowledge-Based Systems* 249 (2022), 108954.
[17] Mehdi Elahi, Dietmar Jannach, Lars Skjærven, Erik Knudsen, Helle Sjovaag, Kristian Tolonen, Oyvind Holmstad, Igor Pipkin, Eivind Throndsen, Agnes Stenbom, Eivind Fiskerud, Adrian Oesch, Loek Vredenberg, and Christoph Trattner. 2021. Towards responsible media recommendation. *AI and Ethics* 2 (2021), 103–114.
[18] Wenqi Fan, Xiangyu Zhao, Xiao Chen, Jingran Su, Jingtong Gao, Lin Wang, Qidong Liu, Yiqi Wang, Han Xu, Lei Chen, and Qing Li. 2022. A comprehensive survey on trustworthy recommender systems. *arXiv preprint arXiv:2209.10117* (2022).
[19] Yingqiang Ge, Shuchang Liu, Zuohui Fu, Juntao Tan, Zelong Li, Shuyuan Xu, Yunqi Li, Yikun Xian, and Yongfeng Zhang. 2022. A survey on trustworthy recommender systems. *arXiv preprint arXiv:2207.12515* (2022).
[20] 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. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 681–691.
[21] Noah Giansiracusa. 2021. Tools for truth. In *How Algorithms Create and Prevent Fake News*. Springer, 217–229.
[22] David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. 1992. Using collaborative filtering to weave an information tapestry. *Communications of the ACM* 35, 12 (1992), 61–70.
[23] Asela Gunawardana and Guy Shani. 2015. Evaluating recommender systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, 265–308.
[24] 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.
[25] Naieme Hazrati and Francesco Ricci. 2022. Recommender systems effect on the evolution of users' choices distribution. *Information Processing & Management* 59, 1 (2022), 102766.
[26] Zengyou He, Xiaofei Xu, Zhexue Joshua Huang, and Shengchun Deng. 2005. FP-outlier: Frequent pattern based outlier detection. *Computer Science and Information Systems* 2, 1 (2005), 103–118.
[27] Keman Huang, Michael Siegel, and Stuart Madnick. 2018. Systematically understanding the cyber attack business: A survey. *ACM Computing Surveys* 51, 4 (2018), 1–36.
[28] Mubashir Imran, Hongzhi Yin, Tong Chen, Quoc Viet Hung Nguyen, Alexander Zhou, and Kai Zheng. 2023. ReFRS: Resource-efficient federated recommender system for dynamic and diversified user preferences. *ACM Transactions on Information Systems* 41, 3 (2023), 1–30.
[29] Dietmar Jannach, Pearl Pu, Francesco Ricci, and Markus Zanker. 2021. Recommender systems: Past, present, future. *AI Magazine* 42, 3 (2021), 3–6.
[30] Dietmar Jannach, Pearl Pu, Francesco Ricci, and Markus Zanker. 2022. Recommender systems: Trends and frontiers. *AI Magazine* 43, 2 (2022), 145–150.
[31] Dietmar Jannach, Paul Resnick, Alexander Tuzhilin, and Markus Zanker. 2016. Recommender systems—Beyond matrix completion. *Communications of the ACM* 59, 11 (2016), 94–102.

[32] Dietmar Jannach and Markus Zanker. 2022. Value and impact of recommender systems. In *Recommender Systems Handbook*. Springer US, 519–546.
[33] Govind Kumar Jha, Manish Gaur, Preetish Ranjan, and Hardeo Kumar Thakur. 2023. A survey on trustworthy model of recommender system. *International Journal of System Assurance Engineering and Management* 14, Suppl. 3 (2023), 789–806.
[34] Di Jin, Luzhi Wang, He Zhang, Yizhen Zheng, Weiping Ding, Feng Xia, and Shirui Pan. 2023. A survey on fairness-aware recommender systems. *arXiv preprint arXiv:2306.00403* (2023).
[35] Joseph A. Konstan and Loren G. Terveen. 2021. Human-centered recommender systems: Origins, advances, challenges, and opportunities. *AI Magazine* 42, 3 (2021), 31–42.
[36] Shyong K. Lam, Dan Frankowski, and John Riedl. 2006. Do you trust your recommendations? An exploration of security and privacy issues in recommender systems. In *Emerging Trends in Information and Communication Security*. Lecture Notes in Computer Science, Vol. 3995. Springer, 14–29.
[37] Jurek Leonhardt, Avishek Anand, and Megha Khosla. 2018. User fairness in recommender systems. In *Companion Proceedings of The Web Conference 2018 (WWW'18)*. 101–102.
[38] Lei Li, Yongfeng Zhang, and Li Chen. 2023. Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems* 41, 4 (2023), 1–26.
[39] Zhiwei Liu, Liangwei Yang, Ziwei Fan, Hao Peng, and Philip S. Yu. 2022. Federated social recommendation with graph neural network. *ACM Transactions on Intelligent Systems and Technology* 13, 4 (2022), 1–24.
[40] Wenpeng Lu, Fanqing Meng, Shoujin Wang, Guoqiang Zhang, Xu Zhang, Antai Ouyang, and Xiaodong Zhang. 2019. Graph-based Chinese word sense disambiguation with multi-knowledge integration. *Computers, Materials & Continua* 61, 1 (2019), 197–212.
[41] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. 2013. Speech enhancement based on deep denoising autoencoder. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH'13)*. 436–440.
[42] Yanzhang Lyu, Hongzhi Yin, Jun Liu, Mengyue Liu, Huan Liu, and Shizhuo Deng. 2021. Reliable recommendation with review-level explanations. In *Proceedings of the 2021 IEEE 37th International Conference on Data Engineering (ICDE'21)*. 1548–1558.
[43] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning disentangled representations for recommendation. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NIPS'19)*. 5711–5722.
[44] Paolo Massa and Paolo Avesani. 2007. Trust-aware recommender systems. In *Proceedings of the 2007 ACM Conference on Recommended Systems (RecSys'07)*. 17–24.
[45] David Massimo and Francesco Ricci. 2023. Combining reinforcement learning and spatial proximity exploration for new user and new POI recommendations. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation, and Personalization (UMAP'23)*. ACM, New York, NY, 164–174.
[46] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54, 6 (2021), Article 115, 35 pages.
[47] Pavel Merinov, David Massimo, and Francesco Ricci. 2022. Sustainability driven recommender systems. In *Proceedings of the 12th Italian Information Retrieval Workshop (IIR'22)*. V1–6.
[48] 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* 7, 4 (2007), 23–es.
[49] Xia Ning, Christian Desrosiers, and George Karypis. 2015. A comprehensive survey of neighborhood-based recommendation methods. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, 37–76.
[50] Cyberspace Administration of China. 2021. Regulation Rules on the Recommendation Algorithm for Internet Information Service. http://politics.people.com.cn/n1/2022/0104/c1001-32323657.html
[51] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. 2021. Deep learning for anomaly detection: A review. *ACM Computing Surveys* 54, 2 (2021), Article 38, 38 pages.
[52] Hossein A. Rahmani, Yashar Deldjoo, Ali Tourani, and Mohammadmehdi Naghiaei. 2022. The unfairness of active users and popularity bias in point-of-interest recommendation. In *Proceedings of the International Workshop on Algorithmic Bias in Search and Recommendation*. 56–68.
[53] Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). 2022. *Recommender Systems Handbook*. Springer.
[54] Xin Rong. 2014. Word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738* (2014).
[55] Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. 2023. Large language models are competitive near cold-start recommenders for language- and item-based preferences. In *Proceedings of the ACM Conference on Recommender Systems (RecSys'23)*. ACM, New York, NY, 890–896.

[56] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on Machine Learning (ICML'16)*. 1670–1679.

[57] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi LI, and Lanyu Xu. 2016. Edge computing: Vision and challenges. *IEEE Internet of Things Journal* 3, 5 (2016), 637–646.

[58] Hyejin Shin, Sungwook Kim, Junbum Shin, and Xiaokui Xiao. 2018. Privacy enhanced matrix factorization for recommendation with local differential privacy. *IEEE Transactions on Knowledge and Data Engineering* 30, 9 (2018), 1770–1782.

[59] Shoujin Wang, Liang Hu, Yan Wang, Longbing Cao, Quan Z. Sheng, and Mehmet Orgun. 2019. Sequential recommender systems: Challenges, progress and prospects. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19)*. 6332–6338.

[60] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurNIPS'20)*. 6827–6839.

[61] Nava Tintarev and Judith Masthoff. 2022. Beyond explaining single item recommendations. In *Recommender Systems Handbook*. Springer US, 711–756.

[62] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008), 2579–2605.

[63] Rongyao Wang, Shoujin Wang, Wenpeng Lu, and Xueping Peng. 2022. News recommendation via multi-interest news sequence modelling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, Los Alamitos, CA, 7942–7946.

[64] Shoujin Wang, Longbing Cao, Yan Wang, Quan Z. Sheng, Mehmet A. Orgun, and Defu Lian. 2021. A survey on session-based recommender systems. *ACM Computing Surveys* 54, 7 (2021), 1–38.

[65] Shoujin Wang, Liang Hu, and Longbing Cao. 2017. Perceiving the next choice with comprehensive transaction embeddings for online recommendation. In *Machine Learning and Knowledge Discovery in Databases:*. Lecture Notes in Computer Science, Vol. 10535. Springer, 285–302.

[66] Shoujin Wang, Liang Hu, Yan Wang, Quan Z. Sheng, Mehmet Orgun, and Longbing Cao. 2020. Intention nets: Psychology-inspired user choice behavior modeling for next-basket prediction. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. 6259–6266.

[67] Shoujin Wang, Liang Hu, Yan Wang, Quan Z. Sheng, Mehmet Orgun, and Longbing Cao. 2020. Intention2Basket: A neural intention-driven approach for dynamic next-basket planning. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*. 2333–2339.

[68] Shoujin Wang, Ninghao Liu, Xiuzhen Zhang, Yan Wang, Francesco Ricci, and Bamshad Mobasher. 2022. Data science and artificial intelligence for responsible recommendations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4904–4905.

[69] Shoujin Wang, Yan Wang, Fikret Sivrikaya, Sahin Albayrak, and Vito Walter Anelli. 2023. Data science for next-generation recommender systems. *International Journal of Data Science and Analytics* 16, 2 (2023), 135–145.

[70] Shoujin Wang, Xiaofei Xu, Xiuzhen Zhang, Yan Wang, and Wenzhuo Song. 2022. Veracity-aware and event-driven personalized news recommendation for fake news mitigation. In *Proceedings of the ACM Web Conference 2022 (WWW'22)*. 3673–3684.

[71] Yan Wang, Lei Li, and Guanfeng Liu. 2015. Social context-aware trust inference for trust enhancement in social network based recommendations on service providers. *World Wide Web* 18, 1 (2015), 159–184.

[72] Yan Wang and Fu-Ren Lin. 2006. Trust and risk evaluation of transactions with different amounts in peer-to-peer e-commerce environments. In *Proceedings of the 2006 IEEE International Conference on e-Business Engineering (ICEBE'06)*. 102–109.

[73] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems* 41, 3 (2023), 1–43.

[74] Yan Wang, Duncan S. Wong, Kwei Jay Lin, and Vijay Varadharajan. 2008. Evaluating transaction trust and risk levels in peer-to-peer e-commerce environments. *Information Systems and e-Business Management* 6, 1 (2008), 25–48.

[75] Zhenlei Wang, Jingsen Zhang, Honteng Xu, Xu Chen, Yongfeng Zhang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Counterfactual data-augmented sequential recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Review (SIGIR'21)*. 347–356.

[76] Le Wu, Xiangnan He, Xiang Wang, Kun Zhang, and Meng Wang. 2023. A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation. *IEEE Transactions on Knowledge and Data Engineering* 35, 5 (2023), 4425–4445.

[77] Zhiang Wu, Junjie Wu, Jie Cao, and Dacheng Tao. 2012. HySAD: A semi-supervised hybrid shilling attack detector for trustworthy product recommendation. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12)*. 985–993.

[78] 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 11th ACM Conference on Recommender Systems (RecSys'17)*. 107–115.

[79] Yuwen Xiong, Mengye Ren, and RaquelUrtasun. 2020. LoCo: Local contrastive representation learning. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurNIPS'20)*. 11142–11153.

[80] Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu. 2019. *Federated Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer.

[81] Haibin Zhang, Yan Wang, and Xiuzhen Zhang. 2012. A trust vector approach to transaction context-aware trust evaluation in e-commerce and e-service environments. In *Proceedings of the 2012 5th IEEE International Conference on Service-Oriented Computing and Applications (SOCA'12)*. 1–8.

[82] Haibin Zhang, Yan Wang, Xiuzhen Zhang, and Ee-Peng Lim. 2015. ReputationPro: The efficient approaches to contextual transaction trust computation in e-commerce environments. *ACM Transactions on the Web* 9, 1 (2015), Article 2, 49 pages.

[83] He Zhang, Bang Wu, Xingliang Yuan, Shirui Pan, Hanghang Tong, and Jian Pei. 2022. Trustworthy graph neural networks: Aspects, methods and trends. *arXiv preprint arXiv:2205.07424* (2022).

[84] Shuai Zhang, Yi Tay, Lina Yao, Aixin Sun, and Ce Zhang. 2022. Deep learning for recommender systems. In *Recommender Systems Handbook*. Springer US, 173–210.

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

[86] Xiuzhen Zhang, Lishan Cui, and Yan Wang. 2013. CommTrust: Computing multi-dimensional trust by mining e-commerce feedback comments. *IEEE Transactions on Knowledge and Data Engineering* 26, 7 (2013), 1631–1643.

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

[88] Xiaoming Zheng, Yan Wang, Mehmet Orgun, Youliang Zhong, and Guanfeng Liu. 2014. Trust prediction with propagation and similarity regularization. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI'14)*. 237–243.

Received 3 July 2023; revised 2 October 2023; accepted 3 October 2023