Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Towards Explainable and Fair Recommender Systems: A Survey of Methods,Taxonomies, and Challenges

Mohammad Reza Hassanpour Charmchi
Shahrood University of Technology
Masoume Gholizade
Semnan University
Shib Sankar Sana
shib_sankar@yahoo.com

Kishore Bharati Bhagini Nivedita College(Co-Ed)

Research Article

Keywords: Recommender Systems, Explainability, Fairness, Trustworthy AI, Bias Mitigation, Interpretable Machine Learning, Survey

Posted Date: June 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9560078/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Towards Explainable and Fair Recommender Systems: A Survey of Methods, Taxonomies, and Challenges

Mohammad Reza Hassanpour Charmchi[a], Masoume Gholizade[b] and Shib Sankar Sana[c,*]

[a]Department of Computer Engineering and IT, Shahrood University of Technology, Shahrood, Iran
[b]Department of Electrical and Computer Engineering, Semnan University, Semnan, Iran
[c]Kishore Bharati Bhagini Nivedita College (Co-Ed), Behala, Kolkata, 700060, India

ARTICLE INFO

Keywords:
Recommender Systems
Explainability
Fairness
Trustworthy AI
Bias Mitigation
Interpretable Machine Learning
Survey

ABSTRACT

Recommender systems have become a fundamental component of modern online platforms, significantly influencing user experiences by delivering personalized content. However, the increasing complexity of contemporary recommendation models has raised critical concerns regarding their transparency and fairness. This survey provides a comprehensive and unified review of explainable and fair recommender systems, emphasizing the need to move beyond accuracy-centric design toward trustworthy and responsible recommendation paradigms. We first introduce the core concepts of recommender systems, explainability, and fairness, and discuss their roles in building transparent and accountable systems. We then propose a multi-dimensional taxonomy that systematically categorizes existing approaches based on objectives, intervention stages, explanation types, and fairness perspectives. Furthermore, we present a detailed analysis of state-of-the-art methods, including explainable recommendation techniques, fairness-aware models, and emerging approaches that jointly address both aspects. In addition, we examine commonly used evaluation metrics and provide a comprehensive analysis of the trade-offs among accuracy, explainability, and fairness, highlighting the challenges of multi-objective optimization in real-world systems. Finally, we identify key open challenges, including data bias, methodological limitations, evaluation inconsistencies, and deployment constraints, and outline promising directions for future research. This survey aims to provide researchers and practitioners with a unified perspective and to support the development of recommender systems that are not only accurate, but also transparent, fair, and trustworthy.

1. Introduction

Recommender systems have become a cornerstone of modern digital platforms, playing a crucial role in shaping user experiences and driving business value [90]. Major services such as Netflix, Amazon, and Spotify rely heavily on recommendation algorithms to personalize content, enhance user engagement, and increase revenue [62]. By filtering large-scale information and presenting relevant items, these systems significantly influence user decisions across domains such as e-commerce, entertainment, and social media [32]. As digital ecosystems continue to expand, the economic and societal impact of recommender systems has grown substantially [46].

Despite their remarkable success, traditional recommender systems have predominantly focused on optimizing predictive accuracy, often at the expense of other critical aspects such as transparency and fairness [69, 15]. The increasing adoption of complex machine learning models, particularly deep learning-based approaches, has further intensified this issue. While these models can achieve high predictive performance, they typically operate as black-box systems, offering limited insight into how recommendations are generated [55]. This lack of interpretability can undermine user trust, hinder system adoption, and restrict the ability of developers to debug, validate, and improve recommendation models.

To address these limitations, explainability has emerged as a fundamental requirement in modern recommender systems [6]. Explainable recommender systems aim to provide understandable and transparent justifications for their recommendations, thereby enhancing user trust, satisfaction, and acceptance [59]. For example, explanations such as "Because you watched similar movies" or "Users with similar preferences liked this item" help users better understand the rationale behind recommendations [1, 14]. Beyond improving user experience, explainability also facilitates system debugging, model validation, and compliance with emerging regulatory frameworks that require transparency in

*Corresponding author

Table 1
A comparative overview of representative studies on explainability and fairness in recommender systems and related domains.

Study	Year	Explainability	Fairness	Joint	Taxonomy	Trade-off
[22]	2021	×	✓	×	✓	✓
[10]	2021	✓	✓	×	×	✓
[38]	2023	×	✓	×	✓	✓
[9]	2023	✓	×	×	×	✓
[9]	2023	✓	×	×	×	✓
[75]	2024	×	✓	×	✓	✓
[60]	2024	✓	✓	✓	×	✓
[61]	2025	✓	✓	✓	×	✓
[77]	2025	✓	×	×	×	✓
[58]	2025	×	✓	×	✓	✓
[55]	2025	✓	×	×	✓	×
[70]	2026	×	×	×	✓	×
[44]	2026	✓	×	×	✓	×
[24]	2026	×	✓	×	×	✓
[19]	2026	×	✓	×	✓	✓
This Work	2026	✓	✓	✓	✓	✓

algorithmic decision-making [87]. In addition, explanations can reveal whether recommendations are influenced by biased patterns such as popularity effects or exposure imbalance.

In parallel, fairness has gained increasing attention as a key requirement for building responsible and inclusive recommender systems [72]. Biases embedded in data or learned by algorithms can lead to unfair outcomes, such as unequal exposure of items or discriminatory treatment of users [5]. Common forms of bias include user bias, item exposure bias, and popularity bias, where frequently interacted or popular items are disproportionately recommended. Such biases can reduce content diversity, reinforce existing inequalities, and negatively affect both users and content providers. Moreover, fairness in recommender systems is inherently multi-stakeholder, involving users, item providers, and platform objectives, which makes its definition and evaluation particularly challenging [51]. Ensuring fairness is therefore essential for maintaining ethical standards and long-term system sustainability [81].

Importantly, explainability and fairness are deeply interconnected rather than independent objectives. Explainability can serve as a powerful tool to uncover and diagnose hidden biases within recommendation models, making fairness issues more transparent and actionable. At the same time, fairness-aware interventions may affect the interpretability of models, introducing additional complexity in system design. Moreover, there exist inherent trade-offs among accuracy, fairness, and explainability, where improving one dimension may come at the expense of others. This interplay highlights the necessity of studying these aspects jointly rather than in isolation.

Although a growing body of literature has explored explainability and fairness in recommender systems, most existing surveys focus on only one of these aspects. As a result, there is a lack of comprehensive studies that provide a unified taxonomy and systematically analyze methods addressing both explainability and fairness. Furthermore, limited attention has been given to approaches that jointly optimize these properties and to the trade-offs involved in balancing them. This fragmentation not only limits a holistic understanding of trustworthy recommender systems, but also hinders the development of unified evaluation frameworks and the systematic design of models that jointly consider accuracy, fairness, and explainability.

To systematically highlight these limitations, Table 1 provides a comparative overview of representative survey papers in terms of the aspects they cover, including explainability, fairness, joint consideration, taxonomy, and trade-off analysis.

Here, "Joint" indicates whether the survey simultaneously addresses both explainability and fairness, while "Trade-off" refers to the analysis of interactions between accuracy, fairness, and explainability.

As shown in Table 1, most existing surveys focus on either explainability or fairness in isolation, while only a limited number of works consider both aspects jointly. Moreover, comprehensive taxonomies and analyses of trade-offs remain underexplored.

To bridge this gap, this paper presents a comprehensive survey of explainable and fair recommender systems. Our survey distinguishes itself from prior work by jointly addressing explainability and fairness within a unified and multi-dimensional framework, explicitly linking methodological design, evaluation metrics, and trade-off analysis across these dimensions. We provide a structured perspective that integrates both dimensions and supports a holistic understanding of trustworthy recommender systems.

The main contributions of this work are as follows:

- We propose a unified and multi-dimensional taxonomy that systematically organizes existing methods across explainability and fairness objectives, intervention stages, and methodological perspectives.
- We systematically review and analyze state-of-the-art approaches in explainable recommendation, fairness-aware recommendation, and joint explainability–fairness optimization.
- We examine evaluation metrics and analyze the trade-offs among accuracy, fairness, and explainability, highlighting key challenges and design considerations.
- We identify open research challenges, including data-related issues, methodological limitations, and evaluation gaps, and outline promising directions for future research.

The rest of this paper is organized as follows. Section 2 describes the survey methodology used in this study, while Section 3 introduces the fundamental concepts of recommender systems, explainability, and fairness. Section 4 presents a comprehensive taxonomy of explainability and fairness from multiple perspectives. Section 5 reviews existing methods, including explainability, fairness-aware, and joint approaches. Section 6 discusses evaluation metrics and analyzes the trade-offs among accuracy, explainability, and fairness. Section 7 outlines open challenges and future research directions. Finally, Section 8 concludes the paper.

2. Survey Methodology

This survey follows a structured literature review methodology to ensure systematic coverage, transparency, and reproducibility. The overall process consists of three main stages: search strategy design, study selection, and qualitative analysis. As illustrated in Fig. 1, the study selection procedure follows a PRISMA-inspired workflow, including identification, screening, eligibility assessment, and final inclusion of relevant studies.

Relevant studies were identified using Boolean combinations of keywords related to recommender systems, explainability, and fairness (e.g., “explainable recommendation”, “fairness-aware recommendation”, and “bias in recommender systems”). The search was conducted using two widely recognized academic sources, namely Google Scholar and Scopus, to ensure broad coverage of both journal and conference publications.

No explicit lower bound was imposed on the publication year during the search process in order to avoid excluding potentially relevant foundational studies. However, in practice, the vast majority of the retrieved and selected papers were published between 2019 and April 2026. This reflects the recent and rapidly growing research interest in explainability and fairness within recommender systems.

As shown in Fig. 1, the initial set of records was obtained through database searching and subsequently refined through a multi-stage screening process. First, duplicate records were removed. Next, title and abstract screening was performed to eliminate clearly irrelevant studies. The remaining papers were then subjected to full-text assessment to determine their eligibility based on predefined criteria.

Studies were included if they (i) explicitly addressed explainability or fairness in recommender systems, (ii) presented methodological, algorithmic, or analytical contributions, and (iii) were published in peer-reviewed journals or conference proceedings in English. Studies were excluded if they were non-scholarly (e.g., blog posts or informal reports), tutorial papers, editorials, or not directly relevant to the scope of this survey.

After applying the above selection criteria, a total of 91 studies were retained and included in the final analysis.

The selected papers were systematically analyzed using a unified framework designed to capture key dimensions of the literature. These dimensions include methodological approaches, types of explanations (e.g., intrinsic vs. post-hoc), fairness perspectives (e.g., individual vs. group fairness), and evaluation strategies (e.g., accuracy, diversity, and fairness-related metrics). This structured analysis forms the foundation for the taxonomy and comparative discussion presented in this survey.

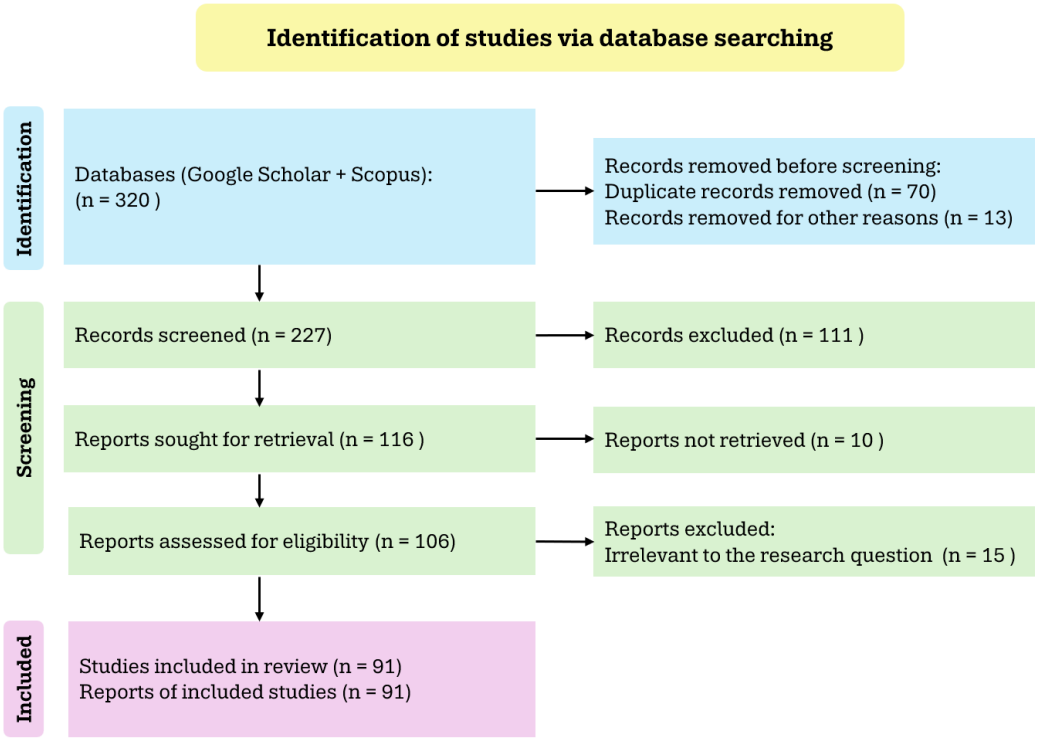


Figure 1: Overview of the literature identification and selection process.

3. Background and Preliminaries

This section provides the necessary background to support the subsequent analysis of explainability and fairness in recommender systems. We begin by introducing the fundamental concepts and modeling paradigms of recommender systems, highlighting their evolution from traditional approaches to modern data-driven and deep learning-based frameworks. We then discuss the concept of explainability, including its objectives, methodological foundations, and its role in enhancing transparency and user trust. Finally, we review the notion of fairness in recommender systems, emphasizing its multi-dimensional nature, stakeholder perspectives, and the challenges associated with defining and measuring fairness. Together, these preliminaries establish a common conceptual foundation for understanding the interplay between accuracy, explainability, and fairness in recommender systems, which will be further explored in the following sections.

3.1. Recommender Systems

Recommender systems are a class of information filtering systems designed to predict users' preferences and provide personalized suggestions [33, 8]. These systems have been widely adopted across domains such as e-commerce, entertainment, and scientific applications to mitigate information overload and support decision-making processes [12].

Formally, a recommender system aims to learn a function $f : U \times I \rightarrow \mathbb{R}$, where U and I denote the sets of users and items, respectively, and the function predicts the preference score of a user $u \in U$ for an item $i \in I$. Depending on the type of available feedback, recommender systems can be categorized into explicit feedback systems (e.g., ratings) and implicit feedback systems (e.g., clicks, views, or purchases), each posing distinct modeling challenges [25].

Traditional recommender systems are generally categorized into collaborative filtering, content-based, and hybrid approaches [63]. Among these, collaborative filtering—particularly matrix factorization—has been extensively utilized due to its capability to model latent user–item interactions effectively [35, 27]. However, such methods rely on latent representations that are inherently difficult to interpret, thereby limiting transparency and explainability for both users and practitioners [76].

To overcome limitations related to data sparsity and restricted feedback, recent research has shifted toward more flexible models capable of handling implicit and sparse data. In this context, graph-based recommender systems have gained significant attention. These approaches represent user–item interactions as graphs and employ techniques such as link prediction to generate recommendations. They are particularly effective in scenarios where explicit feedback or rich metadata is limited, as they can capture higher-order relationships and structural dependencies within the data [35].

Furthermore, modern recommender systems increasingly incorporate additional sources of information to improve representation learning and recommendation quality. Visually-aware recommender systems leverage visual features extracted from item images to better model user preferences, especially in domains where visual attributes are critical [83]. In parallel, semantic-based approaches exploit textual and contextual information using advanced natural language processing techniques, such as BERT and topic modeling, to capture deeper semantic relationships between users and items [54, 13].

In addition, recent advances in deep learning have led to the development of neural recommender systems, including sequential and session-based models that capture temporal dynamics of user behavior. These models, often based on recurrent neural networks or transformer architectures, enable more accurate modeling of user preferences over time but further increase model complexity and reduce interpretability.

In addition to multimodal learning, knowledge graph-based recommender systems have emerged as a powerful paradigm for modeling heterogeneous relational data. By incorporating entities and their relationships into a unified graph structure, these systems can capture rich semantic connections. In particular, graph attention mechanisms enable the model to focus on the most informative relations, thereby improving both recommendation accuracy and interpretability [89].

Beyond model development, recent studies have highlighted the importance of understanding how data characteristics influence recommendation outcomes. Explanatory frameworks based on statistical and regression analysis have been proposed to quantify the impact of various data properties on system performance, including accuracy and fairness. These approaches provide valuable insights into why certain models perform better under specific conditions and contribute to more informed model selection and evaluation [21].

Moreover, recent survey studies indicate that recommender systems are increasingly being applied in complex, data-intensive environments where heterogeneous data sources, uncertainty, and multi-criteria decision-making play a critical role [7]. In such contexts, hybrid and knowledge-based approaches are often preferred due to their ability to integrate diverse information sources and provide more robust recommendations.

Additionally, emerging research directions emphasize the role of recommender systems in broader societal contexts, including sustainability and ethical AI. Recent studies distinguish between designing recommender system to promote sustainable behaviors and developing recommender system that are themselves sustainable in terms of fairness, transparency, and resource efficiency [20]. These perspectives highlight the expanding scope of recommender systems beyond traditional accuracy-oriented objectives.

Despite these advancements, the increasing integration of heterogeneous data sources and sophisticated learning paradigms has led to more complex models. While such complexity often improves recommendation performance, it also introduces significant challenges in terms of transparency, interpretability, and computational cost. Moreover, these challenges are closely related to issues of explainability and fairness, as opaque models may obscure biased decision patterns and hinder accountability.

Consequently, designing recommender systems that achieve a balance between accuracy, explainability, fairness, and efficiency remains an open and critical research challenge.

3.2. Explainability in Recommender Systems

Explainability in recommender systems refers to the ability of a system to provide understandable and meaningful reasons for why a particular item is recommended to a user [86]. In contrast to purely accuracy-driven approaches, explainable recommender systems aim to enhance transparency, interpretability, and user trust by exposing the rationale behind recommendations [4, 66]. However, explainability is not a single, unified concept; rather, it encompasses multiple dimensions related to the purpose of explanations, the knowledge sources used, the generation mechanisms, and the presentation of explanations to users [11, 64].

To address this complexity, Caro-Martínez et al. proposed an ontological formalization that organizes explainability into key components such as motivation, knowledge source, generation process, presentation format, argumentation style, level of detail, and interaction mode [11]. This perspective is particularly important because it distinguishes

between an *explanation* as an artifact presented to users and *explainability* as a broader system property that supports understanding, trust, and decision-making. Moreover, explanations in recommender system can serve different stakeholders, including end users, developers, and regulators, each requiring different forms of transparency and levels of detail [18].

From a functional standpoint, explanations can fulfill multiple objectives beyond simple justification [2]. They can improve user trust, enhance satisfaction, support persuasion, and enable users to scrutinize or refine the recommendation process. Additionally, explanations may rely on diverse types of knowledge, including item attributes, user preferences, contextual signals, and social or collaborative information, highlighting the richness of the explainability design space [11].

One important line of research focuses on semantic or knowledge-based explanations. In this paradigm, recommendations are justified through explicit relationships between items previously preferred by the user and newly recommended items. Musto et al. proposed a framework based on Linked Open Data, where semantic properties extracted from knowledge bases such as DBpedia are used to construct graph-based explanation patterns [56]. These patterns connect user interests with candidate items and can be transformed into natural-language explanations independently of the underlying recommendation algorithm. This characteristic makes the approach particularly suitable for black-box models, where internal decision processes are not directly accessible.

Another prominent direction is aspect-based explanation, which provides fine-grained justification by identifying specific item aspects that match user preferences. Hou et al. introduced an Aspect-based Matrix Factorization (AMF) model that incorporates aspect-level information extracted from user reviews into the recommendation process [34]. Their approach defines quantitative measures such as User Aspect Preference and Item Aspect Quality to capture user interests and item characteristics at a finer granularity. By integrating these signals into the recommendation model, the system not only improves prediction accuracy but also generates more detailed and persuasive explanations grounded in interpretable item features.

Despite these advances, many widely used recommendation models remain inherently difficult to interpret. In particular, latent factor models such as matrix factorization learn hidden user and item representations that are not directly understandable to humans. As highlighted by Vlachos et al., such latent representations limit transparency and motivate the development of intrinsically interpretable recommendation approaches [76].

To address this limitation, several works attempt to incorporate interpretability directly into the model. For instance, Vlachos et al. proposed a co-clustering-based method that identifies overlapping groups of users and items, enabling recommendations to be explained through explicit and human-readable structures [76]. This approach provides a transparent alternative to black-box models while maintaining competitive performance and supporting cold-start scenarios [39].

Beyond model design, recent research emphasizes that explainability must also be systematically evaluated. In practice, assessing explanation quality remains challenging due to the lack of standardized metrics and benchmarks. To address this issue, Coba et al. introduced *RecoXplainer*, a unified framework for developing and evaluating explainable recommender systems [17]. The framework integrates recommendation algorithms, explanation techniques, and evaluation metrics within a complete pipeline, enabling reproducibility and fair comparison across different approaches.

RecoXplainer further highlights that explainability should be treated as a measurable property of recommender systems, evaluated through metrics such as fidelity, explainability precision, and explanation quality. This perspective reinforces the importance of rigorous evaluation methodologies and standardized experimental settings in advancing the field [17].

Another emerging direction focuses on reducing reliance on external knowledge sources while preserving interpretability. Many explainable approaches depend on auxiliary data such as knowledge graphs or textual descriptions, which may not always be available. To address this limitation, Caro-Martínez et al. proposed a graph-based approach that constructs explanations solely from user–item interaction data [12]. By representing interactions as a bipartite graph and applying link prediction techniques, the model generates recommendations that can be explained through observable graph structures such as shared neighbors.

This approach is particularly effective in minimal knowledge scenarios, where explicit metadata is limited. Despite its simplicity, the graph-based representation enables intuitive relational explanations, improving transparency without requiring complex external knowledge sources [12].

More recently, explainable recommender systems have increasingly incorporated structured knowledge and deep learning techniques. Knowledge graph-based models have gained attention due to their ability to capture rich semantic relationships between users, items, and auxiliary entities. In particular, Zhang et al. proposed HKGAT, a heterogeneous

knowledge graph attention network that leverages attention mechanisms to identify informative reasoning paths for recommendation [89]. These paths can be used to generate interpretable explanations that trace how user preferences propagate through the graph.

Such approaches reflect a broader trend toward combining representation learning with symbolic reasoning, enabling systems to achieve both high accuracy and meaningful explanations. By assigning attention weights to different relations, knowledge graph-based models provide a transparent mechanism to highlight the most influential factors in the recommendation process [89].

In parallel with these algorithmic developments, there is a growing recognition that explainability is fundamentally a human-centered property. Traditional systems often adopt a one-size-fits-all approach, providing the same explanation to all users regardless of their needs or expertise. However, recent work by Guesmi et al. emphasizes the importance of interactive and user-controlled explanations [29].

In their approach, explanations are presented at multiple levels of detail, allowing users to progressively explore the reasoning behind recommendations. Users can also interact with explanations through different perspectives, such as *what*, *why*, *how*, and *what-if*, thereby actively shaping their understanding of the system. This interactive paradigm enables a balance between explanation completeness and cognitive load [29].

Empirical findings further indicate that providing controllable and interactive explanations can significantly improve transparency, trust, user satisfaction, and overall user experience. By allowing users to decide how much information they want to access and when, these approaches move beyond static explanations toward adaptive and user-centric explainability [29].

Overall, explainability in recommender systems spans multiple complementary dimensions, including conceptual modeling, semantic and aspect-based explanations, interpretable model design, evaluation frameworks, graph-based methods, and knowledge graph-driven approaches. At the same time, the shift toward human-centered and interactive explanations highlights the importance of aligning technical solutions with user needs. These perspectives provide a comprehensive foundation for the taxonomy and methodological analysis presented in the following sections.

3.3. Fairness in Recommender Systems

Fairness in recommender systems has recently emerged as an important research direction, driven by the increasing societal impact of recommendation technologies [80]. While traditional recommender systems have primarily focused on optimizing predictive accuracy, a growing body of literature suggests that such optimization may lead to biased outcomes that disproportionately disadvantage certain users or item providers. As a result, fairness is increasingly recognized as a key component in the design of trustworthy and responsible recommender systems.

Despite its importance, fairness in recommender system remains difficult to define and operationalize. Unlike accuracy, fairness is inherently multi-dimensional and context-dependent, with no universally accepted definition [84, 85]. Prior studies emphasize that fairness depends on several factors, including the stakeholders involved, the type of benefit being distributed (e.g., relevance, exposure, or utility), and the specific application scenario [5]. Consequently, different fairness notions may coexist, and in some cases, they may even conflict with one another.

A commonly adopted perspective in the literature is to distinguish between different stakeholders affected by recommendation outcomes. Early work primarily focuses on user (consumer) fairness and item or provider fairness. User fairness generally aims to ensure that recommendation quality is equitably distributed across different user groups, particularly those defined by sensitive attributes or activity levels. In contrast, provider fairness focuses on the fair allocation of exposure among items belonging to different providers, aiming to mitigate issues such as popularity bias that favor already dominant items [31]. More recent studies extend this view by considering fairness in a multi-stakeholder setting, where users, providers, and items are simultaneously taken into account, providing a more comprehensive understanding of fairness in recommender systems. [5].

To formalize the diverse notions of fairness, several unified frameworks have been proposed. Among them, the work of Amigó et al. introduces an information-theoretic perspective that models recommendation outcomes as probability distributions and measures fairness using metrics such as Kullback–Leibler divergence and mutual information [5]. This formulation provides a general and flexible approach to capturing different fairness criteria, while also highlighting key dimensions such as the level of comparison (individual versus group), stakeholder perspective, and type of utility being evaluated.

In addition to conceptual definitions, a substantial body of research highlights the role of data characteristics in shaping fairness outcomes. User–item interaction data often exhibit various forms of bias, including imbalanced user activity, skewed rating distributions, and popularity effects. These biases can be amplified by learning algorithms,

resulting in systematic disparities across groups. Empirical analyses suggest that, while data properties can explain variations in recommendation accuracy, fairness is significantly more difficult to model and predict, indicating a complex interplay between data, user behavior, and model design [21].

One of the most widely discussed forms of unfairness in recommender system is exposure bias. In many practical scenarios, frequently interacted or popular items receive disproportionately high visibility, while less popular or minority-group items are underrepresented. This imbalance can reinforce existing inequalities and reduce content diversity. For example, Gómez et al. show that provider groups from dominant regions tend to receive higher exposure compared to those from underrepresented regions, highlighting the impact of data imbalance on recommendation outcomes [31]. To address such issues, many fairness-aware approaches rely on re-ranking strategies that adjust recommendation lists to better balance exposure across groups.

Beyond exposure-based fairness, recent studies explore fairness-aware representation learning techniques. Since modern recommender systems often rely on latent representations, biases present in the data may be encoded into these representations. To mitigate this effect, approaches based on self-supervised learning and adversarial training have been proposed to learn unbiased representations by removing sensitive information from latent spaces. These methods aim to reduce group-based disparities while maintaining recommendation performance [21].

Another important direction in the literature concerns contextual fairness. Several studies argue that fairness should not always be interpreted as strict equality, particularly in real-world applications where different users or items may have different roles or expectations. For instance, in premium service scenarios, users or providers who contribute additional value to the platform may reasonably expect differentiated treatment. In this context, fairness is defined relative to a desired distribution of outcomes rather than uniform allocation. Context-aware fairness frameworks have been proposed to model such scenarios by incorporating application-specific constraints and adjusting recommendation outputs accordingly [31].

Overall, existing research suggests that fairness in recommender systems is a complex and multi-faceted concept that involves multiple stakeholders, data-related biases, and algorithmic considerations. The diversity of definitions and evaluation criteria highlights the need for flexible frameworks and comprehensive methodologies. As recommender systems continue to evolve, balancing fairness with other objectives such as accuracy and explainability remains an open and challenging research problem.

4. Taxonomy of Explainability and Fairness in Recommender Systems

To systematically organize the diverse body of research on explainable and fair recommender systems, this section introduces a unified taxonomy that captures the main conceptual and methodological dimensions of the field. As discussed in previous sections, explainability and fairness are multi-faceted properties that can be analyzed from different perspectives, including system objectives, stages of intervention, explanation mechanisms, and fairness definitions.

Rather than treating these aspects in isolation, the proposed taxonomy provides a structured view that highlights how existing approaches can be categorized along multiple complementary dimensions. Specifically, we organize the literature based on four key perspectives: (i) objective-oriented categorization, which focuses on the primary goals of the methods; (ii) intervention-stage categorization, which considers where fairness or explainability is incorporated in the recommendation pipeline; (iii) explanation-type categorization, which classifies methods based on how explanations are generated and presented; and (iv) fairness-perspective categorization, which reflects different stakeholder and bias considerations.

Fig. 2 presents an overview of the proposed taxonomy, illustrating the relationships between the main dimensions and their corresponding subcategories. This structured representation serves as a conceptual map for the remainder of this section, where each dimension is discussed in detail.

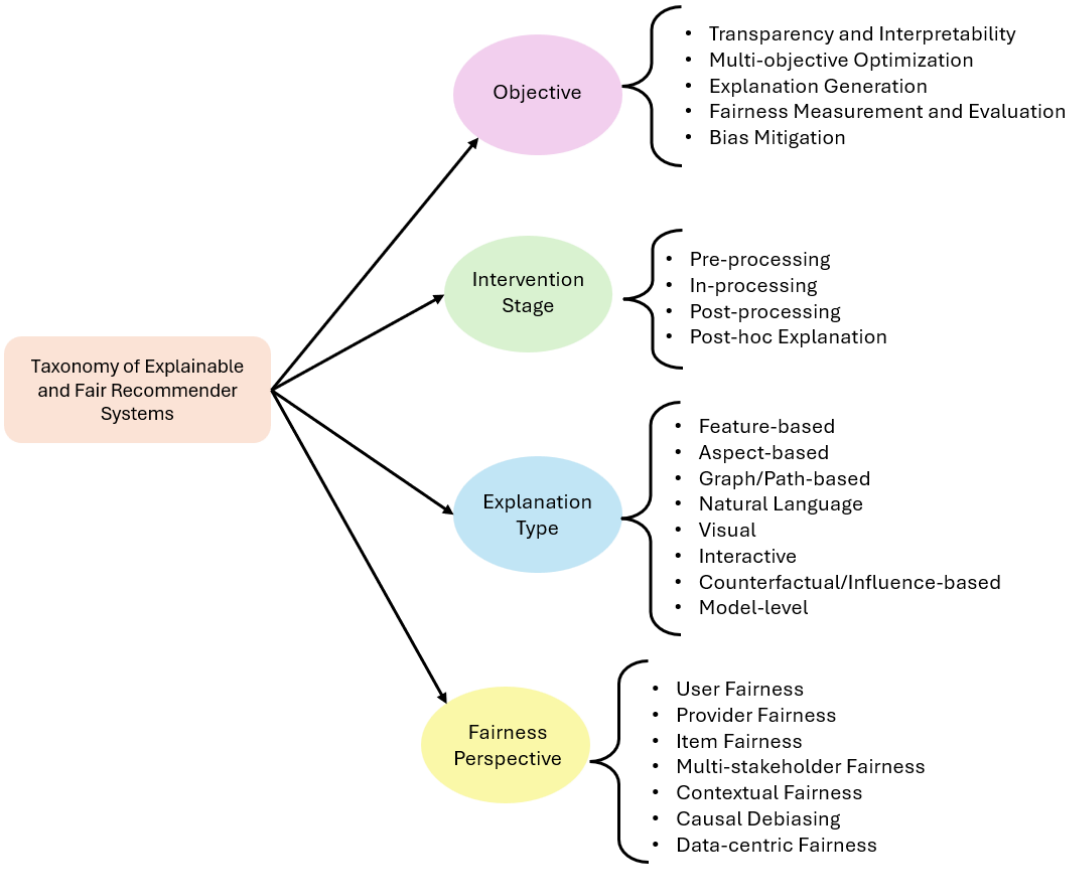


Figure 2: Taxonomy of explainability and fairness in recommender systems.

4.1. Based on Objective

Existing studies on explainable and fair recommender systems can be categorized according to their primary objectives [43]. While these objectives are often interrelated, the literature reveals several dominant directions, including transparency and interpretability, explanation generation, fairness measurement, bias mitigation, and multi-objective optimization.

A first group of studies focuses on *transparency and interpretability*. These works aim to make the internal logic of recommender systems understandable to different stakeholders. Conceptual frameworks highlight that explainability is a multi-dimensional process involving motivation, knowledge sources, generation mechanisms, and presentation strategies, thus extending beyond simple justification generation. Such approaches emphasize that transparency plays a critical role in improving user trust, satisfaction, and decision support.

A second category targets *explanation generation*. These approaches aim to provide explicit and meaningful justifications for recommendations. Recent works move beyond heuristic explanations by incorporating fine-grained semantic information into the recommendation process. For instance, aspect-based models extract features from user reviews and quantitatively model both user preferences and item quality at the aspect level, enabling more persuasive and personalized explanations while improving recommendation accuracy.

Another major objective is *fairness measurement and evaluation*. Traditional recommender systems have primarily focused on accuracy; however, recent research emphasizes the need to evaluate fairness as a complementary dimension. In this context, explanatory frameworks have been proposed to analyze how data characteristics influence both accuracy and fairness. Notably, empirical studies show that data properties can explain a significant portion of accuracy variation, while fairness outcomes are more complex and harder to predict, highlighting the intrinsic difficulty of modeling fairness in recommendation systems [21].

A further line of work focuses on *bias mitigation and fairness-aware recommendation*. These approaches aim to reduce disparities arising from data imbalance, popularity bias, or unequal exposure. A key insight from the literature is that recommender systems operate in multi-stakeholder environments, where fairness must be considered not only for users (consumers) but also for providers and items. Recent studies explicitly distinguish between these stakeholders and demonstrate that fairness should ensure equitable treatment across all of them, including equal opportunity for exposure and balanced recommendation quality [5].

More specifically, fairness-aware approaches can target different stakeholders, including *consumer fairness* (equal recommendation quality across user groups) and *provider fairness* (equal exposure for item providers). For example, provider fairness studies show that imbalances in data—such as geographic dominance of certain content producers—can lead to systematic under-exposure of minority groups. Re-ranking techniques have been proposed to redistribute visibility and exposure across groups while maintaining acceptable recommendation accuracy [31].

Finally, recent works increasingly focus on *multi-objective optimization*, where accuracy, fairness, and sometimes explainability are jointly considered. These approaches recognize that optimizing only for accuracy can lead to unfair outcomes, and therefore introduce optimization frameworks that balance multiple objectives. For instance, fairness-aware frameworks such as CP-Fairness models enable joint optimization of consumer and provider benefits through re-ranking and constrained optimization techniques, demonstrating that fairness can be improved without significantly sacrificing accuracy [57].

Overall, this objective-based taxonomy highlights a clear shift in recommender systems research—from purely accuracy-driven optimization toward broader goals such as transparency, fairness, accountability, and trustworthiness. Moreover, it emphasizes the need to design systems that balance multiple, often competing, objectives in a principled and systematic manner.

4.2. Based on Intervention Stage

Another important dimension for categorizing explainable and fair recommender systems is the stage at which explainability or fairness mechanisms are incorporated into the recommendation pipeline. From this perspective, existing approaches can be broadly classified into four categories: pre-processing, in-processing, post-processing, and post-hoc explanation. Each stage reflects a different level of intervention, ranging from data preparation to model interpretation.

Pre-processing approaches aim to mitigate bias at the data level before model training. These methods typically involve filtering, reweighting, sampling, or restructuring the input data to reduce the impact of skewed interactions and imbalanced distributions. Although relatively fewer studies explicitly focus on this stage, prior work shows that data characteristics—such as sparsity, imbalance, and rating distributions—play a crucial role in shaping both accuracy and

fairness outcomes [21]. Consequently, understanding and addressing data bias is an essential prerequisite for building fair and reliable recommender systems, even before applying algorithmic interventions.

In-processing approaches incorporate fairness or explainability directly into the model training process by modifying the learning objective, representation space, or underlying causal assumptions. In fairness-aware recommendation, a prominent direction is representation-level debiasing, where models aim to learn user–item embeddings that are less sensitive to group-specific or biased signals. For instance, self-supervised and adversarial learning techniques can be used to remove sensitive attributes from latent representations while preserving recommendation performance [49]. Another important line of work focuses on causal debiasing, where methods attempt to disentangle true user preferences from confounding factors such as popularity bias or latent variables through techniques such as instrumental variable learning or multi-view representation modeling [37, 16].

In the context of explainability, in-processing approaches are often associated with intrinsically interpretable models, where explanations are generated as part of the model itself. Examples include knowledge graph-based models that provide reasoning paths, attention-based architectures that highlight influential features, and interpretable latent factor models that expose meaningful structures in user–item interactions [89, 17]. Compared to other stages, in-processing methods offer tighter integration of fairness and explainability, but often at the cost of reduced flexibility and increased model complexity.

Post-processing approaches operate after the recommender system has generated its initial outputs. These methods are particularly popular in fairness-aware recommendation because they allow modifying recommendation lists without altering the underlying model. A common strategy is re-ranking, where items are reordered to achieve a more balanced distribution of exposure across users, providers, or item groups. For example, equity-based re-ranking techniques have been proposed to improve provider fairness by aligning item visibility with group representation [31]. Similarly, model-agnostic frameworks such as PyCPFair optimize fairness for both consumers and producers through post-processing adjustments [57]. More recent approaches extend these techniques to multi-stakeholder settings, balancing fairness across users, providers, and items simultaneously [67]. While post-processing methods are flexible and easy to deploy, they may introduce trade-offs with recommendation accuracy.

Post-hoc explanation approaches focus on interpreting the outputs of recommender systems after they have been generated, typically treating the underlying model as a black box. These methods aim to provide transparency without modifying the internal structure of the recommender. One line of work relies on semantic reasoning, where external knowledge sources such as Linked Open Data are used to construct interpretable connections between user preferences and recommended items, enabling natural language explanations independent of the recommendation algorithm [56]. Another line of work focuses on model diagnostics, including deletion-based analysis and counterfactual sensitivity methods, which evaluate how removing users, items, or interaction patterns affects recommendation outcomes [6, 65]. These approaches provide high-fidelity insights into model behavior and are particularly useful for debugging and auditing, although they can be computationally expensive.

Overall, the intervention-stage taxonomy highlights that explainability and fairness can be incorporated at different points in the recommendation pipeline. Pre-processing focuses on improving the data, in-processing integrates constraints into model training, post-processing adjusts recommendation outputs, and post-hoc methods interpret already generated recommendations. Each stage presents distinct trade-offs in terms of flexibility, scalability, model dependence, and ease of deployment, suggesting that the choice of intervention should depend on application requirements and system constraints.

4.3. Based on Explanation Type

Explainable recommender systems can also be categorized according to the type of explanations they provide. In the literature, explanation types are generally distinguished based on the source of information, the generation mechanism, and the way explanations are presented to users. Common categories include feature-based, aspect-based, graph/path-based, natural language, visual, interactive, counterfactual, and model-level explanations.

Feature-based explanations describe recommendations by highlighting the contribution of input features. Techniques such as LIME and SHAP are widely used to provide local explanations by estimating feature importance [82, 88]. These methods are intuitive and model-agnostic, making them suitable for structured data; however, they may struggle to capture complex feature interactions in highly non-linear models.

Aspect-based explanations focus on item-level aspects (e.g., quality, price, or attributes extracted from user reviews) to justify recommendations. By modeling both user preferences and item characteristics at the aspect level, these approaches provide fine-grained and personalized explanations. For instance, aspect-based matrix factorization

integrates aspect importance into the recommendation process, improving both interpretability and recommendation quality [34].

Graph- and path-based explanations leverage relationships between users, items, and auxiliary entities. In particular, knowledge graph-based approaches explain recommendations through semantic paths that connect user preferences to recommended items [89, 56]. These methods are effective in capturing rich relational information and generating reasoning paths, but their performance and interpretability depend heavily on the completeness and quality of the underlying graph.

Natural language explanations present recommendations in human-readable textual form. Recent approaches employ deep learning models and large language models to generate personalized explanations alongside recommendations [47, 91, 73]. While these explanations are highly persuasive and user-friendly, they raise challenges related to controllability, faithfulness, and objective evaluation.

Visual explanations rely on images or visual features to justify recommendations, particularly in domains such as fashion and multimedia. For example, systems may highlight representative images or visual regions that align with user preferences [59]. These explanations are intuitive but are typically domain-dependent and less applicable to non-visual settings.

Interactive explanations allow users to explore recommendations dynamically by providing multiple levels of detail and user control. Users can refine, query, or navigate explanations based on their needs, leading to improved transparency and user satisfaction. However, such approaches introduce additional design and usability challenges [29].

Counterfactual and influence-based explanations explain recommendations by analyzing how changes in the input data affect model outputs. Techniques such as deletion-based diagnostics evaluate the impact of removing users, items, or interactions on recommendation results [6, 65]. These methods provide high-fidelity insights into model behavior but are often computationally expensive.

Finally, *model-level explanations* are provided by intrinsically interpretable models, where transparency is embedded into the model design. Examples include co-clustering methods and interpretable matrix factorization techniques [76, 35, 11]. While these approaches offer strong interpretability and high fidelity, they may sacrifice flexibility and predictive performance compared to complex black-box models.

Overall, different explanation types provide complementary perspectives on transparency. Their effectiveness depends on factors such as data characteristics, application domain, and user requirements. Moreover, these explanation types differ in their ability to support tasks such as user trust enhancement, system debugging, and bias detection, highlighting the importance of selecting appropriate explanation strategies in practice.

4.4. Based on Fairness Perspective

Fairness-aware recommender systems can be categorized according to the perspective from which fairness is defined and evaluated. Unlike accuracy, fairness is inherently multi-dimensional and context-dependent, and thus different perspectives emphasize different stakeholders, objectives, and sources of bias. The literature generally distinguishes several key perspectives, including user fairness, provider fairness, item fairness, multi-stakeholder fairness, contextual fairness, causal debiasing, and data-centric fairness.

User fairness focuses on whether different user groups receive comparable recommendation quality. In this perspective, protected or group attributes such as gender, age, or activity level are used to examine disparities in ranking performance or utility. For instance, Deldjoo et al. show that fairness variations across user groups are more difficult to explain than accuracy variations [21]. To mitigate such disparities, representation learning approaches—such as self-supervised and adversarial methods—aim to remove group-specific information from latent representations [49].

Provider fairness considers whether item providers receive equitable exposure in recommendation lists. This perspective is particularly relevant in multi-sided platforms where recommendations directly affect the visibility and opportunities of sellers, authors, or content creators. Gómez et al. demonstrate that data imbalance and popularity bias may lead to over-exposure of dominant provider groups [31]. To address this issue, re-ranking strategies and exposure-aware methods have been proposed to better align visibility with fairness objectives [31].

Item fairness shifts the focus from providers to individual items. The goal is to prevent scenarios where a small subset of popular items dominates recommendation lists while long-tail items remain underrepresented. This perspective is closely related to popularity bias and exposure imbalance and is often treated as a distinct fairness dimension in recommendation outcomes [67].

Multi-stakeholder fairness extends fairness considerations to multiple parties simultaneously. In such settings, users, providers, items, and platforms may have competing objectives, making fairness inherently a multi-objective

problem. Frameworks such as PyCPFair enable joint optimization of consumer and provider fairness through model-agnostic re-ranking techniques [57]. Other approaches further incorporate additional stakeholders, emphasizing that fairness should be evaluated across all affected entities [26, 67].

Contextual fairness argues that fairness should be defined relative to the application context rather than as strict equality. In some scenarios, such as premium services, differentiated treatment of users or items may be expected. Context-aware frameworks model fairness as alignment with a desired distribution of outcomes, taking into account business rules, platform policies, and domain-specific constraints [45].

Causal and debiasing perspectives focus on identifying and mitigating the root causes of unfairness rather than only adjusting final outputs. These approaches aim to disentangle true user preferences from confounding factors such as popularity bias, social influence, or selection bias. For example, causal representation learning methods leverage techniques such as instrumental variables and multi-view modeling to remove latent confounders and improve fairness and robustness [37, 16, 78].

Data-centric fairness emphasizes the role of dataset characteristics in shaping fairness outcomes. Instead of treating unfairness solely as an algorithmic issue, this perspective examines how properties such as sparsity, imbalance, skewed activity distributions, and sensitive attributes influence both accuracy and fairness. Information-theoretic frameworks provide a unified way to quantify fairness under different conditions [5], while empirical studies show that fairness is often harder to explain than accuracy due to complex interactions between data and model behavior [21].

Overall, these perspectives highlight that fairness in recommender systems is not a single objective but a collection of complementary and sometimes conflicting criteria. Selecting an appropriate fairness notion depends on the target stakeholders, the type of bias being addressed, and the specific application context. This diversity further underscores the importance of flexible and multi-objective approaches in designing fair recommender systems.

5. Explainability and Fairness Methods in Recommender Systems

This section provides a comprehensive overview of the main methodological approaches developed to improve explainability and fairness in recommender systems. While traditional recommender models primarily focus on predictive accuracy, recent research has introduced a wide range of techniques to enhance transparency, interpretability, and equitable outcomes.

To organize this diverse body of work, we categorize existing methods into three main groups: explainability methods, fairness-aware approaches, and joint explainability–fairness models. This categorization highlights both the methodological differences and the increasing trend toward integrating multiple objectives within unified recommendation frameworks.

5.1. Explainability Methods

Explainability methods in recommender systems aim to make recommendation processes and outputs more understandable to users, domain experts, and system developers. Existing approaches differ in terms of whether explanations are generated as part of the recommendation model itself or produced after recommendations have been generated. Accordingly, explainability methods can be broadly grouped into intrinsically interpretable models, post-hoc explanation techniques, knowledge- and graph-based methods, natural language explanation models, interactive explanation interfaces, and domain-specific explainable recommender systems [53].

A major line of research focuses on *intrinsically interpretable recommender models*, where transparency is embedded into the model design. These methods attempt to avoid purely black-box recommendation mechanisms by introducing interpretable structures, factors, or representations. For example, co-clustering-based matrix factorization methods provide explanations through overlapping user–item groups, enabling recommendations to be associated with meaningful clusters rather than opaque latent factors [76].

In such models, user–item interactions are often formulated through interpretable co-cluster affiliations. Specifically, the probability of interaction between a user u and an item i can be expressed as:

$$P(r_{ui} = 1) = 1 - e^{-\langle f_u, f_i \rangle} \tag{1}$$

where f_u and f_i denote the co-cluster affiliation vectors of users and items, respectively, and $\langle f_u, f_i \rangle$ captures their shared cluster memberships. This formulation provides a direct interpretation of recommendations based on overlapping group structures.

Similarly, physically interpretable recommender systems incorporate domain-specific physical process features into matrix factorization, allowing learned parameters to correspond to meaningful concepts in computational model selection [36]. In healthcare and scientific domains, interpretable deep or causal models have also been used to support decision-making, such as causal Bayesian networks for treatment recommendation [79] and biologically interpretable neural recommender models for drug response prediction [40]. These methods offer high explanation fidelity because the explanation is directly tied to the internal structure of the model, although they are often domain-dependent and may require carefully designed features or expert knowledge.

Another important group consists of *feature- and attribution-based explanation methods*. These approaches explain recommendations by identifying the input features that most strongly influence model predictions. Model-agnostic methods such as LIME and SHAP are widely used in this category because they can explain black-box models without requiring access to internal model parameters. For instance, LIME has been applied to crop recommendation systems to provide local feature-level explanations for individual recommendations [82]. SHAP and LIME have also been combined in clinical recommender systems to provide both local and global explanations for diagnosis and treatment support [23, 88].

Among these methods, SHAP provides a theoretically grounded approach based on cooperative game theory [88]. The contribution of each feature j to a prediction is quantified using Shapley values:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{j\}) - f(S)] \tag{2}$$

where F denotes the set of all features and S represents a subset excluding feature j. This formulation captures the marginal contribution of each feature across all possible feature combinations, enabling both local and global interpretability.

Similar SHAP-based explanation strategies have been used in federated knowledge graph recommender systems [41], mobile application recommendation [74], and other security- or healthcare-oriented decision-support scenarios. These methods are useful for improving transparency and trust, but they may provide limited insight into complex feature interactions and are often evaluated qualitatively rather than through standardized explanation metrics.

Review- and aspect-based explainable recommendation represents another well-established direction. These methods exploit user reviews, item attributes, or textual feedback to generate fine-grained explanations. Aspect-based matrix factorization models integrate user aspect preferences and item aspect quality into the recommendation process, allowing recommendations to be justified in terms of specific item aspects [34]. Neural models further combine reviews, attributes, and attention mechanisms to identify important textual or attribute-level signals behind predictions [52]. Such methods are particularly effective in domains where rich textual reviews are available, as they can generate explanations that are more specific and persuasive than general feature-importance scores. However, their performance depends strongly on the quality of aspect extraction, review availability, and textual representation learning.

Graph-based and knowledge-based methods provide explanations by leveraging relationships among users, items, and auxiliary entities. In minimal-knowledge scenarios, interaction graphs can be used to generate interpretable recommendations through link prediction and neighborhood-based reasoning [12]. Knowledge graph-based methods extend this idea by incorporating semantic entities and relations, enabling recommendations to be explained through reasoning paths. For example, heterogeneous knowledge graph attention networks use attention mechanisms over graph paths to identify influential relations for recommendation [89].

Figure 3 illustrates a representative knowledge graph-based explainable recommender framework, where user preferences are propagated over a heterogeneous graph and aggregated via attention mechanisms to identify informative relational paths.

Linked Open Data-based explanation frameworks similarly construct semantic connections between items previously liked by a user and newly recommended items, producing algorithm-independent explanations [56]. Visually-aware heterogeneous information networks also use meta-path-based reasoning to generate scalable and explainable recommendations in visual domains [54]. These approaches are powerful for relational and semantic explanation, but their quality depends on the completeness, accuracy, and scalability of the underlying graph structure.

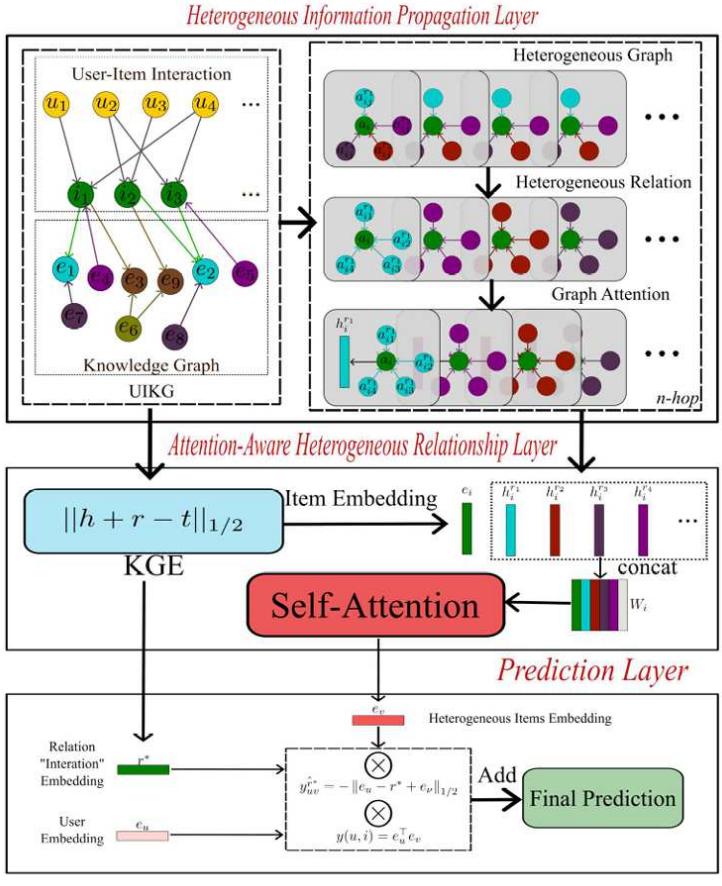


Figure 3: Framework of a heterogeneous knowledge graph attention network (HKGAT), where user preferences are propagated over a knowledge graph and aggregated using attention mechanisms to generate explainable recommendations [89].

Natural language explanation methods have gained increasing attention due to their ability to present recommendations in user-friendly textual form. Early and recent neural approaches jointly model rating prediction and explanation generation, often using encoder–decoder architectures or multi-task learning frameworks [91]. More recent methods employ large language models and pretrained language models to generate personalized explanations for sequential recommendation [47]. Reinforcement learning and simulated human feedback have also been used to optimize explanation quality according to criteria such as informativeness, persuasiveness, and personalization [73]. In reinforcement learning-based recommendation scenarios, large language models can further translate policy decisions into understandable textual explanations [71]. Although natural language explanations improve user understanding and persuasiveness, they raise important challenges related to faithfulness, controllability, hallucination, and objective evaluation.

Interactive explanation methods shift the focus from static explanations to user-controlled explanation processes. Instead of presenting a single fixed explanation, these systems allow users to explore recommendations at different levels of detail and from different perspectives. For example, interactive scientific literature recommendation systems provide explanations based on "what", "why", "how", and "what-if" questions, enabling users to inspect and refine the recommendation rationale [29]. Online explainable recommender systems also integrate explanation generation with online learning, allowing the system to adapt recommendations and explanations based on user feedback over time [48]. These approaches are valuable for improving transparency, trust, and user engagement, but they introduce additional challenges related to interface design, cognitive load, and evaluation complexity.

Post-hoc diagnostic and counterfactual explanation methods aim to explain recommender systems after the recommendation model has already been trained. Unlike user-facing textual or visual explanations, these methods are often designed for developers and auditors. Deletion-based diagnostics measure how removing users, items, or interactions affects recommendation performance, thereby identifying influential data points and model sensitivities [6]. To improve scalability, spectral biclustering-based approaches group users and items into blocks and perform block-level deletion analysis, offering a more tractable form of counterfactual explanation [65]. Such approaches are useful for debugging, robustness analysis, and accountability, but they are computationally expensive and are not always suitable for direct end-user explanation.

Several explainability methods are also tailored to specific domains and modalities. In visual recommendation, Bayesian ranking methods select representative images as explanations, making recommendations more intuitive in domains such as fashion and multimedia [59]. In academic recommendation, semantic models based on BERT and LDA extract explainable keywords to justify research-related recommendations [83]. In legal recommendation, interpretable hand-crafted features are used to model case–statute relations for statute ranking [42]. In clinical recommender systems, explainability is often achieved through causal graphs, feature importance, dashboards, or biologically meaningful latent factors [79, 23, 40, 88]. These domain-specific methods demonstrate that explanation design should be aligned with the needs, knowledge, and decision-making context of target users.

Finally, toolkits and conceptual frameworks have been proposed to support the systematic development and evaluation of explainable recommender systems. Ontological models provide a structured vocabulary for describing explanation components, including explanation goals, knowledge sources, generation mechanisms, and presentation formats [11]. RecoXplainer offers a unified library for developing and evaluating explainable recommender systems, integrating recommendation algorithms, explanation methods, and offline evaluation metrics such as fidelity and explainability precision [17]. These frameworks are important because they address the lack of standardization and reproducibility in explainable recommendation research.

Overall, explainability methods in recommender systems span a broad methodological spectrum, ranging from intrinsically interpretable models and feature-attribution techniques to graph-based reasoning, natural language generation, interactive interfaces, and post-hoc diagnostics. Each method offers different strengths: intrinsic models provide high fidelity, attribution methods are flexible and model-agnostic, graph-based methods support semantic reasoning, natural language methods improve user understanding, and diagnostic methods help developers audit model behavior. However, major challenges remain, including the lack of standardized evaluation protocols, limited user studies, scalability issues, and the difficulty of ensuring that explanations are faithful to the actual recommendation process.

5.2. Fairness-aware Methods

Fairness-aware recommender systems aim to mitigate biases and ensure equitable treatment of different stakeholders, including users, item providers, and platforms. Unlike traditional recommender systems that primarily optimize

accuracy, fairness-aware approaches explicitly incorporate fairness constraints, objectives, or evaluation criteria into the recommendation process. Existing methods can be broadly categorized into fairness measurement frameworks, re-ranking and post-processing approaches, causal and debiasing models, multi-stakeholder fairness methods, and learning-based fairness-aware representations.

A fundamental line of research focuses on *fairness definition and measurement*. Since fairness is inherently context-dependent and multi-dimensional, several works aim to unify its conceptualization and evaluation. For instance, an information-theoretic framework based on Kullback–Leibler divergence provides a generalized formulation that unifies different fairness notions across recommender systems [5]. Such approaches measure fairness as the divergence between the observed and desired benefit distributions:

$$D_{KL}(P_\theta \parallel Q) = \sum_{x \in A_X} P_\theta(x) \log \frac{P_\theta(x)}{Q(x)} \tag{3}$$

where P_θ denotes the observed distribution of exposure, utility, or effectiveness across groups, and Q represents the target fair distribution. This formulation allows different fairness criteria, such as parity or proportionality, to be expressed within a unified framework.

Similarly, data-centric analyses show that fairness outcomes are strongly influenced by data characteristics such as sparsity, imbalance, and user-item distributions, often making fairness more difficult to explain and control than accuracy [21]. These studies highlight that fairness is not a single objective but a family of metrics that must be carefully selected depending on the application context and stakeholders involved.

Another major category includes *re-ranking and post-processing fairness methods*, which adjust recommendation outputs to improve fairness without modifying the underlying model. These approaches are widely used due to their flexibility and model-agnostic nature. For example, provider fairness can be achieved by re-ranking items to align exposure with user preference distributions while minimizing accuracy loss [28]. Equity-based re-ranking methods further ensure fair visibility across different provider groups, such as geographic regions, by reducing exposure disparities [31]. Optimization-based frameworks such as PyCPFair extend this idea to multi-stakeholder settings by jointly optimizing consumer and provider fairness using mathematical programming techniques [57]. Although these approaches are effective and easy to integrate, they typically operate as post-processing steps and may not fully address biases embedded in learned representations.

Context-aware and constrained fairness methods extend traditional fairness models by considering real-world constraints and contextual factors. In premium recommendation scenarios, fairness may depend on contextual variables such as user subscription level or item status, requiring different fairness distributions rather than strict equality [45]. Similarly, in environments with limited item availability, fairness-aware allocation models explicitly consider supply constraints and optimize exposure across multiple stakeholders using inter- and intra-group fairness metrics [68]. These approaches emphasize that fairness must be adapted to practical constraints and domain-specific requirements, rather than being defined as a uniform objective.

A rapidly growing direction involves *causal and debiasing-based fairness methods*, which aim to identify and remove bias at its source rather than correcting it after recommendation. Causal disentanglement techniques separate user preferences from confounding factors such as social influence, enabling more accurate and less biased recommendations [78]. Instrumental variable-based representation learning further addresses latent confounding by learning surrogate variables that isolate causal effects from spurious correlations [37]. Multi-view causal representation learning methods extend this idea by leveraging multiple data sources to mitigate confounding, selection, and popularity biases simultaneously [16]. These approaches provide strong theoretical grounding for fairness but often require complex modeling assumptions and may lack interpretability for end users.

Another important class includes *learning-based fairness-aware models*, which incorporate fairness objectives directly into the training process. For example, self-supervised learning combined with adversarial debiasing can learn unbiased user representations by removing sensitive attribute information such as gender, thereby improving both fairness and accuracy [50]. These models typically optimize a joint objective that balances recommendation performance with fairness constraints, making them more robust than post-hoc methods. However, they often require access to protected attributes and involve complex training procedures.

Multi-stakeholder fairness has also gained significant attention, as recommender systems often impact multiple parties simultaneously. Hypergraph-based ranking models address this challenge by modeling higher-order relationships among users, items, and providers, enabling fair exposure for under-represented stakeholders while maintaining

recommendation quality [26]. More comprehensive frameworks explicitly distinguish between consumer, provider, and item fairness, and optimize them jointly using entropy-based or exposure-based objectives [67]. These approaches reflect the growing recognition that fairness must be considered from a holistic perspective rather than focusing solely on users or items.

Finally, several studies focus on *bias analysis and fairness from behavioral and systemic perspectives*. Research on cognitive and behavioral bias demonstrates that recommender systems can amplify phenomena such as confirmation bias, especially when recommendations are presented as conclusive decisions [16]. Similarly, analyses of social and interaction data reveal that factors such as hyperactive users or popularity dynamics can significantly distort recommendation outcomes [31]. These works highlight that fairness issues are not only algorithmic but also influenced by user behavior, system design, and data generation processes.

Overall, fairness-aware methods in recommender systems encompass a wide range of techniques, from metric design and post-processing adjustments to causal modeling and fairness-aware learning. Each category offers distinct advantages: re-ranking methods are flexible and easy to deploy, causal models address bias at its root, and learning-based approaches integrate fairness into model training. However, key challenges remain, including the trade-off between fairness and accuracy, the need for context-aware and multi-stakeholder fairness definitions, limited availability of protected attributes, and the lack of standardized evaluation protocols. Future research is expected to focus on integrating fairness with explainability, developing dynamic and context-aware fairness models, and enabling real-world deployment of fair recommender systems.

5.3. Joint Explainability and Fairness Approaches

Joint explainability and fairness approaches aim to address transparency and equity simultaneously in recommender systems. Compared with studies that focus exclusively on either explainability or fairness, this research direction is still relatively limited. Most existing works do not optimize explainability and fairness as two explicit objectives at the same time; rather, they combine fairness-aware learning or debiasing with interpretable mechanisms, or use explainability as a tool to analyze fairness-related behavior.

One important direction is *explainable debiasing*, where interpretable or semi-interpretable representations are used to separate useful preference signals from biased signals. For example, causal disentanglement methods separate user interest from social influence in social recommendation, helping regulate social influence bias while improving recommendation performance [78]. Similarly, self-supervised and adversarial learning approaches aim to remove group-specific information from learned representations, thereby improving user fairness while providing a representation-level view of how bias is mitigated [50]. Although these methods are not always user-facing, they contribute to model transparency by making the debiasing process more structured and analyzable.

Another group of studies combines fairness-aware recommendation with multi-stakeholder modeling. Hypergraph-based ranking methods model high-order relationships among users, items, and providers, allowing the system to improve the coverage of under-represented stakeholders while preserving recommendation utility [26]. Model-agnostic re-ranking frameworks such as PyCPFair further support joint optimization of consumer and producer fairness, making fairness interventions explicit and easier to inspect [57]. More recent multi-stakeholder frameworks extend this perspective by distinguishing consumers, providers, and items as separate fairness dimensions [67]. These approaches are particularly relevant to joint explainability–fairness research because they make fairness objectives visible at the ranking or allocation level.

Post-hoc analytical approaches also play an important role in connecting explainability and fairness. Data-centric analysis can explain variations in both accuracy and fairness by relating model behavior to dataset characteristics such as sparsity, imbalance, and sensitive attributes [21]. Such methods do not directly mitigate unfairness, but they provide diagnostic insight into why fairness disparities arise. In this sense, explainability serves as an auditing mechanism for fairness.

Recent work has also explored broader ethical recommender systems, where explainability is treated as a prerequisite for fairness, accountability, and trustworthiness [30]. In addition, explainable reciprocal recommender systems incorporate transparency mechanisms such as SHAP-based explanations together with diversity- or fairness-oriented ranking components, illustrating how explainability and bias mitigation can be integrated in practical two-sided recommendation scenarios [3].LLM- and reinforcement learning-based explainable recommendation methods also introduce multi-objective optimization strategies for balancing recommendation quality and explanation quality [73]. Although fairness is not always explicitly measured, such frameworks provide a foundation for future joint optimization of fairness, accuracy, and explainability.

These approaches typically formulate explanation generation as a reinforcement learning problem, where model parameters are updated based on feedback signals that reflect explanation quality. Specifically, policy-based optimization can be expressed as:

$$\theta_{t+1} = \theta_t + \alpha \nabla \log \pi_\theta(a|s)\, r \tag{4}$$

where $\pi_\theta(a|s)$ denotes the policy for generating explanations given state s, α is the learning rate, and r represents reward signals derived from human or simulated feedback [73]. This formulation enables the model to iteratively improve explanation quality by reinforcing actions that lead to more informative and persuasive explanations.

Overall, joint explainability and fairness in recommender systems remains an emerging research area. Current studies mainly connect the two objectives through debiasing, fairness-aware re-ranking, multi-stakeholder modeling, and post-hoc fairness diagnostics. A key limitation is that many methods either provide explainability without formal fairness guarantees or improve fairness without user-facing explanations. Future research should therefore develop unified models that jointly optimize accuracy, fairness, and explainability, provide faithful explanations for fairness interventions, and evaluate these objectives under consistent multi-stakeholder benchmarks.

To facilitate a structured comparison of the discussed methods, Table 2 summarizes representative approaches across explainability, fairness-aware, and joint modeling paradigms. The table highlights the diversity of methodological designs, ranging from intrinsically interpretable models and post-hoc explanation techniques to graph-based reasoning, natural language generation, and fairness-aware optimization frameworks.

As shown in Table 2, explainability methods primarily focus on improving transparency through feature attribution, semantic reasoning, and human-understandable explanations. In contrast, fairness-aware approaches aim to regulate exposure, representation, and bias across different stakeholders. Joint approaches attempt to bridge these objectives, often through debiasing strategies or multi-stakeholder modeling; however, they remain relatively limited compared to standalone explainability or fairness methods.

Table 2
Summary of representative explainability and fairness methods

Reference	Method	Type	Explanation / Fairness Mechanism
[76]	Overlapping co-clustering for interpretable matrix factorization (OCuLaR)	Explainability (Intrinsic)	User–item co-clusters as interpretable latent factors
[79]	Causal Bayesian Network for clinical recommendation	Explainability (Intrinsic)	Graph-based causal reasoning between clinical variables
[36]	Physically interpretable matrix factorization	Explainability (Intrinsic)	Latent factors mapped to physical process features
[82]	LIME-based local explanations for recommendation	Explainability (Post-hoc)	Local surrogate models for feature importance
[23, 88]	SHAP-based explainable recommender systems	Explainability (Post-hoc)	Shapley value-based feature attribution (local + global)
[34]	Aspect-based matrix factorization (AMF)	Explainability (Intrinsic)	User preferences and item quality modeled via aspects
[89]	Heterogeneous Knowledge Graph Attention Network (HKGAT)	Explainability (Graph-based)	Attention over knowledge graph paths for reasoning
[56]	Linked Open Data-based explanation framework	Explainability (Post-hoc)	Semantic relations between items via LOD graphs
[47]	LLM-based sequential recommendation	Explainability (Intrinsic + Natural Language)	Textual explanations generated via pretrained language models
[73]	RL + LLM-based explanation optimization	Explainability (Joint generation)	Policy learning with reward signals for explanation quality
[29]	Interactive explainable recommender system	Explainability (Interactive)	User-controlled explanations (what, why, how, what-if)
[6, 65]	Post-hoc diagnostic explanation methods	Explainability (Post-hoc diagnostic)	Deletion-based and counterfactual analysis
[5]	Information-theoretic fairness framework	Fairness (Measurement)	KL-divergence and mutual information for fairness quantification
[28, 31]	Provider fairness via re-ranking	Fairness (Post-processing)	Exposure balancing across provider groups
[57]	PyCPFair framework	Fairness (Optimization-based)	Joint consumer-provider fairness optimization
[45]	Context-aware fairness	Fairness (Contextual)	Fairness defined based on contextual distributions
[68]	Fair allocation under limited availability	Fairness (Multi-stakeholder)	Exposure optimization under supply constraints
[78]	Causal disentanglement for debiasing	Joint (Explainability + Fairness)	Separation of user interest and social influence
[50]	Self-supervised adversarial debiasing	Fairness (Learning-based)	Removing sensitive attributes from representations
[3]	Explainable reciprocal recommender system	Joint (Explainability + Fairness)	SHAP explanations combined with diversity-aware ranking
[26]	Hypergraph-based multi-stakeholder fairness	Fairness (Graph-based)	Modeling high-order relations for fair exposure
[21]	Data-centric fairness analysis	Joint (Analysis)	Statistical analysis of fairness vs data characteristics

These methodological developments motivate the need for comprehensive evaluation frameworks that can jointly assess accuracy, explainability, and fairness, which is discussed in the next section.

6. Evaluation Metrics and Trade-offs

Evaluation plays a central role in assessing the effectiveness, transparency, and social responsibility of recommender systems. Traditional recommender-system evaluation has mainly focused on predictive accuracy, measuring how well a model estimates user preferences or ranks relevant items. However, in explainable and fairness-aware recommender systems, accuracy alone is insufficient, because a model may achieve high predictive performance while remaining opaque, biased, or unsuitable for real-world deployment.

Therefore, this section reviews the main evaluation dimensions used in the literature, including accuracy metrics, explainability metrics, fairness metrics, and multi-objective trade-off analysis. In particular, it discusses how different metrics capture different aspects of recommender-system quality and why these dimensions should be considered jointly rather than independently.

6.1. Accuracy Metrics

Accuracy evaluation remains a central component in recommender-system assessment because it quantifies how effectively a model predicts user preferences, estimates interaction likelihoods, or ranks relevant items in recommendation lists [44]. However, the notion of accuracy is not uniform across recommender-system tasks. Rating prediction, Top-N ranking, sequential recommendation, clinical decision support, and reinforcement learning-based recommendation each require different evaluation criteria. Therefore, accuracy metrics should be selected according to the output structure of the recommender, the type of user feedback, and the practical decision context.

For rating prediction tasks, the most common metrics are Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These metrics compare the predicted rating $\hat{r}_{ui}$ with the observed rating r_{ui} for a user–item pair (u, i). MAE measures the average absolute deviation between predicted and actual values, while RMSE penalizes larger prediction errors more strongly due to the squared loss term. They can be formally expressed as:

$$MAE = \frac{1}{|\mathcal{T}|} \sum_{(u,i)\in\mathcal{T}} |r_{ui} - \hat{r}_{ui}|, \tag{5}$$

$$RMSE = \sqrt{\frac{1}{|\mathcal{T}|} \sum_{(u,i)\in\mathcal{T}} (r_{ui} - \hat{r}_{ui})^2}, \tag{6}$$

where $\mathcal{T}$ denotes the test set. These metrics are particularly suitable for explicit-feedback recommender systems and have been widely used in aspect-based and review-based explainable recommendation models [34]. Similar error-based measures are also used in language-model-based explainable recommendation, where rating prediction and explanation generation are optimized jointly [91].

In classification-oriented recommender systems, especially in clinical, diagnostic, and decision-support scenarios, accuracy is often evaluated using metrics such as Accuracy, Precision, Recall, F1-score, Area Under the Curve (AUC), and Brier Score. Precision measures the proportion of recommended or predicted positive cases that are actually relevant, whereas Recall measures the proportion of relevant cases that are successfully retrieved. The F1-score combines Precision and Recall through their harmonic mean:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{7}$$

Such metrics are suitable when the recommendation output corresponds to a class label, risk category, treatment option, or diagnostic decision. For example, clinical treatment recommender systems based on causal Bayesian networks use AUC and Brier Score to evaluate discrimination and calibration in survival-related treatment recommendation [79]. Similarly, explainable cancer-treatment recommendation models employ F1-score to assess the correctness of multi-level clinical recommendations [23].

For Top-N recommendation, ranking-based metrics are more appropriate than rating prediction metrics because users are typically exposed only to the highest-ranked items. Precision@K measures the proportion of relevant items

among the top-K recommendations, while Recall@K measures the fraction of all relevant items that appear in the top-K list:

$$Precision@K = \frac{|Rel_u \cap Rec_u^K|}{K}, \tag{8}$$

$$Recall@K = \frac{|Rel_u \cap Rec_u^K|}{|Rel_u|}, \tag{9}$$

where Rel_u is the set of relevant items for user u and Rec_u^K is the set of top-K recommended items. Recall-based evaluation is particularly common in implicit-feedback and one-class collaborative filtering settings, where the main objective is to retrieve relevant items from a large candidate space [76].

Normalized Discounted Cumulative Gain (NDCG) is another widely used ranking metric because it accounts not only for whether relevant items are recommended, but also for their positions in the ranked list. It assigns higher importance to relevant items appearing at earlier ranks:

$$DCG@K = \sum_{j=1}^{K} \frac{2^{rel_j} - 1}{\log_2(j+1)}, \tag{10}$$

$$NDCG@K = \frac{DCG@K}{IDCG@K}, \tag{11}$$

where rel_j denotes the relevance score of the item at position j, and $IDCG@K$ is the ideal DCG value. NDCG is especially important in sequential and knowledge graph-based recommender systems, where correctly ranking the most relevant next item or semantically related item is central to system performance [47]. Knowledge graph-based explainable recommender models also commonly report NDCG and Recall to evaluate ranking quality while simultaneously producing path-based explanations [89].

Mean Average Precision (MAP) evaluates ranking performance by averaging precision values at the ranks where relevant items occur. It is particularly useful when multiple relevant items exist and the order of their retrieval matters:

$$AP_u = \frac{1}{|Rel_u|} \sum_{k=1}^{K} Precision@k \cdot \mathbb{I}(i_k \in Rel_u), \tag{12}$$

$$MAP = \frac{1}{|U|} \sum_{u \in U} AP_u. \tag{13}$$

MAP has been used in reciprocal and two-sided recommender systems, where ranking quality must be evaluated for matching scenarios involving multiple stakeholders [3]. In such systems, ranking accuracy is not only a matter of user preference prediction but also affects exposure, matching quality, and downstream fairness.

Hit Ratio (HR) is also frequently used in Top-N and sequential recommendation. HR@K indicates whether at least one relevant item appears in the top-K list:

$$HR@K = \frac{1}{|U|} \sum_{u \in U} \mathbb{I}(Rel_u \cap Rec_u^K \neq \emptyset). \tag{14}$$

Although HR is intuitive and easy to interpret, it provides less information than NDCG or MAP because it does not fully account for the number or position of relevant items. Nevertheless, it remains useful in next-item prediction tasks, where the goal is often to determine whether the ground-truth next item appears in the recommendation list [37].

In online and reinforcement learning-based recommender systems, accuracy cannot be fully captured by static offline metrics. Instead, evaluation often includes cumulative reward, regret, and long-term interaction utility. Cumulative reward measures the total benefit collected by the recommendation policy over time:

$$R_T = \sum_{t=1}^{T} r_t, \tag{15}$$

where r_t denotes the reward observed at interaction step t. Regret compares the performance of the learned policy with an optimal policy:

$$Regret(T) = \sum_{t=1}^{T} (r_t^* - r_t), \tag{16}$$

where r_t^* is the reward of the optimal action at time t. These metrics are particularly relevant for online explainable recommender systems, where the model must balance recommendation accuracy, exploration, exploitation, and explanation feedback over time [48]. In reinforcement learning-based recommendation with large language model explanations, collection rate, reward, and text-generation quality may be evaluated together to assess both recommendation effectiveness and explanation usefulness [71].

Accuracy evaluation also depends strongly on the experimental protocol. Random train–test splitting is common in traditional rating prediction, but temporal splitting is more appropriate in sequential recommendation, news recommendation, premium recommendation, and real-world deployment scenarios because it better reflects future-item prediction [45]. In fairness-aware and debiasing studies, biased-training and unbiased-testing protocols are sometimes used to examine whether a model generalizes beyond biased historical interactions [37]. This distinction is important because a recommender can achieve high accuracy under random splits while still reproducing popularity bias, exposure bias, or confounding effects.

Overall, accuracy metrics provide essential but incomplete evidence of recommender-system quality. MAE and RMSE are appropriate for explicit rating prediction, Precision@K, Recall@K, MAP, HR, and NDCG are suitable for ranked recommendation, and reward-based metrics are necessary for online or reinforcement learning-based systems. However, high accuracy does not necessarily imply interpretability, fairness, robustness, or user trust. Therefore, modern recommender-system evaluation must consider accuracy together with explainability and fairness metrics rather than treating predictive performance as the sole optimization target.

6.2. Explainability Metrics

Evaluating explainability in recommender systems is inherently challenging due to its subjective, context-dependent, and multi-dimensional nature. Unlike accuracy, which can be quantified through well-defined numerical metrics, explainability involves both objective properties of explanations and subjective human perception. As a result, existing evaluation approaches can be broadly categorized into intrinsic (model-oriented) metrics, proxy metrics, and human-centered evaluation.

Intrinsic metrics aim to quantify how well explanations reflect the internal behavior of the recommendation model. One of the most important criteria is *fidelity*, which measures the degree to which an explanation faithfully approximates the true decision-making process of the underlying model. High-fidelity explanations ensure that the explanation is not merely plausible but accurately represents the model's reasoning. Another important concept is *explainability precision*, which evaluates whether the features, aspects, or relations highlighted in the explanation are truly relevant to the recommendation outcome [17]. In addition, consistency and stability are often considered, ensuring that similar inputs produce similar explanations and that explanations remain robust under small perturbations.

For feature attribution methods such as SHAP and LIME, evaluation focuses on the stability, consistency, and local faithfulness of feature importance scores. Stability refers to whether similar instances yield similar attribution patterns, while local fidelity measures how well the explanation approximates the model in the neighborhood of a specific instance. These properties are particularly important in high-stakes domains such as healthcare and agriculture, where feature-level explanations must be reliable and interpretable [82]. However, such methods may suffer from sensitivity to input perturbations and approximation errors when explaining highly non-linear models.

In graph-based and knowledge-based recommender systems, explainability is typically evaluated through the quality of reasoning paths, relational coherence, and semantic interpretability. Path-based explanations should reflect

meaningful connections between users and items, and their quality depends on the relevance and interpretability of the underlying graph structure. Metrics in this setting often assess whether extracted paths correspond to plausible semantic relationships and whether attention mechanisms correctly identify influential nodes and relations [89]. In Linked Open Data-based approaches, explanation quality is further evaluated based on the correctness and usefulness of semantic connections derived from external knowledge sources [56].

Natural language explanation models introduce additional evaluation challenges due to the generative nature of explanations. Automatic text generation metrics such as BLEU, ROUGE, and BERTScore are commonly used to measure lexical similarity, fluency, and semantic overlap between generated explanations and reference texts [47]. However, these metrics primarily capture surface-level similarity and may fail to assess whether the explanation is faithful to the model or genuinely informative. Recent approaches attempt to address this limitation by incorporating multi-perspective evaluation criteria, including informativeness, persuasiveness, and personalization, often optimized through reinforcement learning frameworks [73].

Human-centered evaluation remains a crucial component of explainability assessment, as the ultimate goal of explanations is to improve user understanding, trust, and decision-making. User studies typically measure perceived transparency, trustworthiness, satisfaction, and cognitive effort. These evaluations often involve controlled experiments, questionnaires, or interactive systems where users assess explanation usefulness in real decision-making scenarios [29]. In addition, qualitative evaluation frameworks examine how explanations influence user behavior, including reliance on recommendations and susceptibility to cognitive biases.

Interactive explanation systems introduce further evaluation dimensions, such as usability, controllability, and cognitive load. Providing more detailed explanations can improve transparency but may also overwhelm users, leading to a trade-off between explanation completeness and user effort. Evaluating such systems therefore requires considering both system performance and human factors, including user expertise and task complexity.

Despite significant progress, a major limitation in explainability evaluation is the lack of standardized and universally accepted metrics. Different studies adopt heterogeneous evaluation protocols, datasets, and criteria, making direct comparison across methods difficult. This fragmentation highlights the need for unified benchmarking frameworks and reproducible evaluation pipelines that can systematically assess explainability alongside other objectives such as accuracy and fairness.

6.3. Fairness Metrics

Fairness evaluation in recommender systems aims to quantify disparities in exposure, utility, or representation across different users, items, or stakeholders. Unlike accuracy, fairness is not a single well-defined objective but a family of context-dependent criteria that depend on application requirements, stakeholder priorities, and societal considerations. As a result, fairness metrics must be carefully designed to reflect the specific notion of fairness relevant to the recommendation scenario.

A general and theoretically grounded approach to fairness measurement is based on information-theoretic formulations. In this framework, fairness is quantified as the divergence between the observed distribution of recommendation outcomes and a predefined target distribution. Measures such as Kullback–Leibler (KL) divergence and mutual information are commonly used to capture discrepancies between these distributions. This formulation provides a unified perspective that can express multiple fairness notions, including statistical parity, proportional fairness, and exposure balance within a single mathematical framework [5].

Group fairness metrics are among the most widely used approaches in recommender systems. These metrics evaluate disparities across predefined sensitive groups, such as gender, age, or geographic region. A central concept in this setting is exposure fairness, which measures whether items or users from different groups receive comparable visibility in recommendation lists. Typical metrics include exposure disparity, defined as the difference in exposure between groups, and visibility-based measures such as ΔExposure, which quantify deviations from a fair allocation of recommendation opportunities [31]. In provider-oriented scenarios, fairness is often assessed by comparing the distribution of recommended items across groups to their expected or desired proportions in the catalog [28].

Beyond group-level analysis, individual fairness focuses on ensuring that similar users receive similar recommendations. This notion is often formalized through similarity-based constraints, where the difference between recommendation outputs should be bounded for users with similar preferences or profiles. However, measuring individual fairness is challenging in practice due to the difficulty of defining appropriate similarity functions in high-dimensional spaces. Recent approaches address this issue through causal fairness frameworks, which aim to eliminate the influence of confounding variables and spurious correlations on recommendation outcomes. Techniques such

as causal disentanglement separate intrinsic user preferences from external influences, enabling a more principled evaluation of fairness through bias reduction [78]. Similarly, instrumental variable-based methods attempt to isolate causal effects by leveraging surrogate variables, allowing fairness to be evaluated through counterfactual consistency and robustness to latent confounding [37].

In multi-stakeholder recommender systems, fairness extends beyond users to include item providers and platform-level objectives. In such settings, fairness metrics must simultaneously account for multiple dimensions, including consumer fairness (fair treatment of users), provider fairness (fair exposure of items or content creators), and item fairness (balanced representation of items). Metrics such as coverage quantify the diversity of recommended items, while entropy-based measures capture the uniformity of exposure distributions across stakeholders. Allocation-based disparity metrics further evaluate how recommendation opportunities are distributed under constraints such as limited item availability or competition among providers [26]. More comprehensive frameworks explicitly model fairness across multiple stakeholders by combining exposure-based and distributional metrics into unified optimization objectives [67]. In constrained environments, fairness is often evaluated using inter-group and intra-group disparity measures that reflect both global and local fairness conditions [68].

Another important class of fairness metrics focuses on bias quantification. These metrics measure the extent to which recommendation outcomes are influenced by biases such as popularity bias, exposure bias, or activity bias. For example, popularity bias can be evaluated by measuring the over-representation of highly popular items in recommendation lists relative to their relevance. Similarly, exposure bias metrics assess whether certain groups systematically receive more visibility than others, regardless of their relevance or quality. These bias-oriented metrics are often used in conjunction with fairness-aware learning and re-ranking approaches to evaluate the effectiveness of debiasing techniques.

Despite the diversity of existing metrics, fairness evaluation remains fundamentally challenging. Different fairness definitions may conflict with each other, and optimizing one fairness criterion can lead to unintended consequences for another. Moreover, fairness metrics are highly sensitive to data characteristics such as sparsity, imbalance, and user behavior patterns, making it difficult to generalize results across datasets and domains. These challenges highlight the need for standardized evaluation protocols and context-aware fairness definitions that can better capture the complexity of real-world recommender systems.

To consolidate the discussion of accuracy, explainability, and fairness metrics, Table 3 summarizes the evaluation practices reported in representative studies. The table shows how different works operationalize evaluation through accuracy metrics, explanation-oriented criteria, fairness measures, and experimental settings.

Table 3
Evaluation metrics used in explainable and fair recommender systems.

Year / Ref.	Evaluation focus	Accuracy metrics	Explainability / fairness metrics	Evaluation setting
2019 [76]	Accuracy–interpretability	Recall@M	Co-cluster interpretability	Offline; industrial + public datasets
2019 [34]	Rating + explanation	RMSE, MAE	Aspect-level explanation	Offline; review datasets
2020 [52]	Review-based explainability	RMSE	Attention-based explanation	Offline; Amazon datasets
2021 [21]	Fairness vs accuracy	nDCG, MAP, Precision	MAD, GCE, bias disparity	Offline simulations
2021 [26]	Multi-stakeholder fairness	Precision	Coverage; provider fairness	Time-aware evaluation
2021 [40]	Clinical interpretability	Pearson, RMSE	Biological interpretation	Offline evaluation
2021 [48]	Online recommendation	Reward, regret	Qualitative explanation	Online + offline
2022 [45]	Contextual fairness	NDCG@5, HR@5	Contextual fairness	Temporal split
2022 [31]	Provider fairness	NDCG	ΔExposure	Offline
2022 [57]	Multi-stakeholder fairness	Precision, Recall, nDCG	GAP fairness	Offline framework
2022 [17]	Explainability benchmarking	Precision, nDCG	Fidelity, MEP	Offline toolkit
2022 [50]	Fair representation	NDCG@20	Fairness gap	Offline
2022 [30]	Ethical evaluation	N/A	Transparency, fairness, accountability	Conceptual
2022 [83]	Semantic explainability	Accuracy	Keyword importance	Offline
2023 [12]	Graph explainability	Recall	Graph interpretability	Cross-validation
2023 [36]	Interpretable MF	Accuracy	Physical interpretability	Offline expert data
2023 [29]	Interactive explanations	N/A	Trust, usability, cognitive load	User study
2023 [91]	Text explanation	RMSE	BLEU, ROUGE	Offline
2023 [54]	Visual explainability	Top-N accuracy	Meta-path explainability score	Offline
2023 [56]	Knowledge-based explanation	N/A	Trust, transparency	User study
2024 [59]	Visual explanation ranking	NDCG	Explanation ranking quality	Offline
2025 [79]	Clinical prediction	AUC, Accuracy, Brier	Causal interpretability	Offline clinical
2025 [41]	Federated XAI	RMSE, MAE	SHAP explanations	Offline
2025 [71]	RL + LLM explanation	Reward, BLEU, ROUGE	Expert evaluation	Offline + expert
2025 [78]	Causal debiasing	NDCG, Recall	Bias regulation	Offline
2025 [23]	Clinical XAI	F1-score	SHAP explanation	Offline
2025 [47]	Sequential RS	HR@K, NDCG@K	BLEU, ROUGE	Offline
2025 [28]	Fair re-ranking	NDCG, Recall	Disparity metrics	Offline
2025 [82]	Local explainability	N/A	LIME feature importance	Offline
2025 [73]	LLM + RL XAI	RMSE, MAE	BLEU, ROUGE, BERTScore	Offline
2025 [89]	KG explainability	NDCG@K, Recall@K	Path-based explanation	Offline
2025 [6]	Post-hoc diagnostics	MAP, Precision, Recall	Influence scores	Retraining protocol
2025 [88]	Clinical SHAP+LIME	Accuracy, AUC	SHAP + LIME	Offline
2026 [3]	Reciprocal RS	NDCG, MAP, Recall	SHAP, diversity@K	Offline
2026 [37]	IV debiasing	Recall@K, NDCG@K	IOU debiasing	Offline
2026 [68]	Limited-item fairness	N/A	Fairness disparity	Offline
2026 [16]	Multi-view debiasing	NDCG@K	Bias reduction	Offline
2026 [65]	Scalable XAI	Ranking sensitivity	Counterfactual explanation	Offline
2026 [67]	Multi-stakeholder fairness	NDCG	Entropy-based fairness	Offline

As shown in Table 3, evaluation practices vary considerably across recommender-system tasks. Accuracy-oriented studies mainly rely on rating prediction or ranking metrics, whereas explainability-oriented studies use fidelity, attribution quality, path interpretability, text-generation metrics, or user studies. Fairness-aware works, in contrast, focus on exposure, visibility, representation, and stakeholder-level disparity. This diversity confirms that evaluation in modern recommender systems must be aligned with the specific objective and deployment context of the model.

6.4. Multi-objective Trade-off Analysis

Modern recommender systems are inherently multi-objective, requiring the simultaneous optimization of accuracy, explainability, and fairness. While traditional systems primarily focus on maximizing predictive performance, recent advances highlight that improving one objective often leads to the degradation of others. This section provides a unified analysis of these trade-offs and proposes a conceptual perspective for understanding their interactions.

A fundamental trade-off exists between accuracy and explainability. Highly accurate models, particularly deep learning-based and graph neural network approaches, often operate as black-box systems with limited interpretability. In contrast, intrinsically interpretable models, such as co-clustering or feature-based approaches, provide transparent reasoning but may sacrifice predictive performance due to reduced model complexity. This tension reflects the classical accuracy–interpretability dilemma, where increased model expressiveness typically reduces transparency [76].

Similarly, the relationship between accuracy and fairness introduces another critical trade-off. Fairness-aware methods, especially re-ranking and exposure-balancing techniques, often modify recommendation outputs to reduce disparities across groups. While these methods effectively improve fairness metrics, they may alter the original ranking and lead to a measurable decrease in accuracy metrics such as NDCG or Recall [31]. More advanced approaches attempt to mitigate this issue by jointly optimizing fairness and relevance through constrained or multi-objective optimization frameworks, although perfect alignment between the two objectives remains difficult [57].

The interaction between explainability and fairness is more nuanced and less explored. On one hand, explainability can serve as a diagnostic tool to identify and understand biases in recommender systems, enabling fairness auditing and transparency in decision-making processes [21]. On the other hand, certain explanation methods may unintentionally reinforce biases by highlighting spurious or popularity-driven features, thereby affecting perceived fairness. Conversely, fairness-aware models that rely on complex debiasing or causal representations may reduce transparency for end users, as their internal mechanisms are not always easily interpretable [78].

Recent research has begun to address these challenges through multi-objective optimization frameworks. For example, reinforcement learning-based recommender systems model recommendation and explanation generation as a policy optimization problem, where multiple reward signals—such as accuracy, explanation quality, and user satisfaction—are jointly optimized [73]. Similarly, multi-stakeholder fairness models incorporate constraints or regularization terms to balance competing objectives across users, providers, and platforms [67]. These approaches often rely on Pareto optimality, where no objective can be improved without degrading another, providing a principled way to analyze trade-offs in high-dimensional objective spaces.

From a methodological perspective, trade-off analysis can be categorized into three main strategies. First, *post-processing approaches* adjust recommendation outputs (e.g., via re-ranking) to satisfy fairness or diversity constraints while preserving as much accuracy as possible. Second, *joint learning approaches* incorporate multiple objectives directly into the training process through multi-task learning or regularized optimization. Third, *interactive and human-in-the-loop approaches* allow users to dynamically control trade-offs, for example by selecting between more accurate or more diverse recommendations.

Despite these advances, several open challenges remain. Existing studies often evaluate trade-offs using limited metrics and do not provide a unified framework for comparing different objectives. Moreover, trade-offs are highly context-dependent and vary across domains, datasets, and user populations. There is also a lack of standardized benchmarks for jointly evaluating accuracy, explainability, and fairness, which hinders reproducibility and real-world deployment.

To address these gaps, this survey highlights the need for a unified multi-objective evaluation framework that explicitly models the interactions among accuracy, explainability, and fairness. Such a framework should support consistent metric definitions, enable Pareto-based comparison of methods, and incorporate user-centric evaluation criteria. Developing such integrated evaluation methodologies represents a key direction for future research in explainable and fair recommender systems.

To further clarify these interactions, Tables 4 and 5 summarize the main multi-objective trade-offs identified across the reviewed studies. These tables focus on how prior work balances competing objectives such as accuracy, explainability, fairness, diversity, scalability, privacy, and user effort.

Table 4
Multi-objective trade-offs in explainable and fair recommender systems (Part 1).

Year / Ref.	Objectives	Trade-off type	Key observation	Implication
2021 [48]	Accuracy + Adaptivity	Exploration vs Exploitation	Increasing exploration improves long-term reward but may reduce short-term accuracy	Online systems must balance learning and performance
2021 [21]	Accuracy + Fairness	Fairness vs Accuracy	Fairness varies significantly across data distributions and may conflict with ranking quality	Data characteristics strongly influence fairness outcomes
2021 [26]	Precision + Fairness	Coverage vs Precision	Increasing exposure of minority providers reduces precision slightly	Multi-stakeholder fairness requires controlled trade-offs
2022 [45]	Accuracy + Fairness	Context-dependent trade-off	Fairness is not uniform across contexts (e.g., premium vs non-premium users)	Fairness must be defined relative to context
2022 [31]	Accuracy + Fairness	Exposure vs Accuracy	Improving provider visibility introduces minor ranking degradation	Re-ranking is effective for practical fairness
2022 [57]	Accuracy + Fairness	Multi-objective optimization	Joint optimization can balance consumer and provider fairness with minimal accuracy loss	Optimization frameworks enable flexible trade-offs
2022 [50]	Accuracy + Fairness	Representation vs Utility	Removing group bias improves fairness but may alter ranking performance	Representation learning impacts fairness-utility balance
2023 [36]	Accuracy + Interpretability	Complexity vs Overfitting	Increasing model complexity improves accuracy but harms interpretability and generalization	Simpler interpretable models remain valuable
2023 [29]	Explainability + Usability	Detail vs Cognitive Load	More detailed explanations improve transparency but increase user effort	Human-centered design must balance clarity and simplicity
2023 [91]	Accuracy + Explainability	Accuracy vs Text Quality	Generating natural language explanations introduces noise in prediction optimization	Joint learning requires careful balancing
2023 [54]	Accuracy + Explainability	Explainability vs Scalability	Rich explainable structures increase computational cost	Efficient architectures are required for large-scale systems
2024 [59]	Accuracy + Explainability	Quality vs Efficiency	High-quality explanation ranking improves user experience but increases computational overhead	Efficient explanation selection is critical
2024 [30]	Ethics + Transparency + Fairness	Multi-objective conflict	Ethical objectives cannot be optimized simultaneously	Multi-objective frameworks are necessary
2025 [79]	Accuracy + Interpretability	Accuracy vs Interpretability	Causal models maintain interpretability with strong predictive performance	Suitable for high-stakes domains (e.g., healthcare)
2025 [41]	Accuracy + Privacy + Explainability	Privacy vs Performance	Federated learning preserves privacy but may affect model performance	Trade-offs emerge in distributed environments

Table 5
Multi-objective trade-offs in explainable and fair recommender systems (Part 2).

Year / Ref.	Objectives	Trade-off type	Key observation	Implication
2025 [71]	Reward + Explainability	Reward vs Explanation Quality	RL policies optimize reward but explanation quality depends on prompt design	Explanation quality is sensitive to model design
2025 [78]	Accuracy + Debiasing	Accuracy vs Bias Reduction	Removing social bias improves fairness but may alter recommendation structure	Bias mitigation is inherently trade-off-driven
2025 [23]	Accuracy + Transparency	Performance vs Transparency	High explainability improves trust without major accuracy loss	Important for clinical applications
2025 [47]	Accuracy + Explainability	Joint optimization	LLM-based models can generate explanations without degrading accuracy	Promising direction for unified modeling
2025 [28]	Accuracy + Fairness	Fairness vs Accuracy	Fairness-aware re-ranking improves exposure equity with minimal accuracy loss	Practical deployment feasible
2025 [73]	Accuracy + Explainability	Pareto trade-off	Multiple explanation objectives require Pareto optimization	Advanced optimization methods needed
2025 [89]	Accuracy + Interpretability	Accuracy vs Reasoning Transparency	Knowledge graph reasoning improves explainability but depends on graph quality	Data quality is critical for explainability
2025 [6]	Fidelity + Scalability	Fidelity vs Computational Cost	Retraining-based explanations are accurate but computationally expensive	Approximation methods are needed
2025 [88]	Accuracy + Interpretability	Accuracy vs Transparency	SHAP and LIME improve interpretability but increase model complexity	Trade-off between simplicity and insight
2026 [3]	Accuracy + Diversity + Explainability	Multi-objective trade-off	Increasing diversity improves fairness but may reduce precision	Balanced ranking strategies required
2026 [37]	Accuracy + Debiasing	Accuracy vs Confounding Bias	Removing confounders improves causal validity but changes ranking outcomes	Causal modeling is essential
2026 [68]	Fairness + Relevance	Fairness vs Utility	Limited item availability intensifies fairness trade-offs	Resource constraints amplify trade-offs
2026 [16]	Accuracy + Debiasing	Accuracy vs Latent Bias	Multi-view learning reduces hidden bias at the cost of higher complexity	More robust but computationally intensive
2026 [65]	Explainability + Scalability	Fidelity vs Scalability	Block-level explanations scale better but lose granularity	Approximate explanations are necessary
2026 [67]	Fairness (Multi-stakeholder)	Fairness vs Accuracy	Balancing consumers, providers, and items reduces overall utility	Requires careful weighting of stakeholders

Table 5 continues the summary of multi-objective trade-offs across more recent studies.

A detailed examination of Tables 4 and 5 shows that trade-offs are not marginal implementation issues, but core characteristics of modern recommender systems. The most frequent conflict is between accuracy and fairness, especially in re-ranking, allocation, and exposure-balancing methods where improving stakeholder fairness may slightly reduce ranking utility. Another recurring conflict appears between accuracy and explainability, where transparent or explanation-oriented models must balance predictive performance with interpretability.

The tables also show that recent studies increasingly move beyond pairwise trade-offs toward multi-objective settings. LLM-based, reinforcement learning-based, causal debiasing, and multi-stakeholder models introduce additional tensions involving explanation quality, computational cost, scalability, privacy, diversity, and user effort. Therefore, accuracy, explainability, and fairness should not be evaluated independently. Instead, they should be treated as interacting objectives within a unified evaluation framework, ideally supported by standardized benchmarks and Pareto-aware comparison protocols.

7. Open Challenges and Future Directions

Although explainability and fairness have received increasing attention in recommender systems, the reviewed literature shows that the field remains fragmented and several important challenges are still unresolved. These challenges extend beyond algorithmic design and involve data quality, conceptual ambiguity, evaluation inconsistency, and real-world deployment constraints.

A major challenge concerns the biased and incomplete nature of recommendation data. User–item interaction data are inherently influenced by exposure mechanisms, popularity effects, and historical system behavior, rather than representing unbiased user preferences. Prior studies show that such data characteristics significantly affect both accuracy and fairness outcomes, often amplifying disparities across user and provider groups [21, 31]. This feedback loop not only leads to unfair exposure but also affects explainability, since generated explanations may reflect biased interaction patterns rather than genuine user interests.

Another critical issue is the limited availability of sensitive attributes and high-quality side information. Many fairness-aware approaches rely on demographic or group-level attributes, which are often unavailable or restricted due to privacy concerns. Similarly, explainable recommender systems frequently depend on auxiliary information such as reviews, knowledge graphs, or textual descriptions, which may be incomplete or domain-dependent [12, 89]. These limitations highlight the need for data-efficient, privacy-aware, and domain-agnostic approaches to both fairness and explainability.

From a methodological perspective, a key challenge is the lack of unified frameworks that jointly optimize accuracy, fairness, and explainability. Existing methods typically focus on one or two of these objectives in isolation. While fairness-aware models attempt to reduce disparities and explainable models improve transparency, their integration remains limited. Moreover, explainability can help reveal bias, but may also produce misleading justifications if explanations are not faithful to the model behavior [17]. Conversely, fairness-aware techniques such as re-ranking or representation debiasing may increase model complexity and reduce interpretability.

The trade-off among these objectives remains one of the most challenging issues. Improving fairness often requires modifying ranking outputs or learned representations, which can degrade recommendation accuracy [31, 57]. Similarly, intrinsically interpretable models may sacrifice predictive performance compared to deep black-box models [76]. Recent approaches attempt to address this problem using multi-objective optimization frameworks, but achieving an optimal balance remains context-dependent and application-specific.

Another important open issue is the faithfulness and reliability of explanations. Many explanation techniques generate plausible and user-friendly justifications, but these explanations are not always faithful to the underlying decision-making process. This issue is particularly evident in post-hoc and natural language explanation methods, where explanations may be persuasive but not causally grounded [17, 73]. Ensuring that explanations accurately reflect model behavior is essential for building trustworthy systems.

Fairness definition also remains an open challenge. As highlighted in prior work, fairness is inherently multi-dimensional and depends on stakeholders, context, and the type of utility being distributed [5]. Different fairness notions, such as user fairness, provider fairness, and multi-stakeholder fairness, may conflict with each other, making it difficult to define a universally acceptable fairness objective. This complexity requires context-aware and flexible fairness frameworks.

Evaluation is another major limitation in current research. While accuracy metrics are relatively standardized, explainability and fairness metrics vary significantly across studies, making comparisons difficult. Explainability is often evaluated using proxy measures or limited user studies, while fairness is assessed using context-specific metrics [21]. This lack of standardization hinders reproducibility and benchmarking. Developing unified evaluation frameworks that jointly assess accuracy, fairness, and explainability is therefore an important direction for future work.

Furthermore, most existing studies rely on static offline evaluation settings, which fail to capture the dynamic nature of real-world recommender systems. User preferences, item distributions, and platform policies evolve over time, and fairness or explainability interventions may have long-term effects through feedback loops [45]. Future research should consider temporal dynamics, longitudinal evaluation, and online experimentation.

Scalability and deployment also present significant challenges. Many fairness-aware and explainable methods require additional computation, complex optimization, or access to large auxiliary data sources. For example, knowledge graph-based explanations and causal debiasing methods can be computationally expensive and difficult to deploy at scale [89, 37]. Developing lightweight and scalable solutions remains essential for real-world applications.

Finally, privacy and human-centered considerations introduce additional complexity. Fairness evaluation may require sensitive attributes, while explainability often depends on user data, raising privacy concerns. At the same time, different stakeholders require different forms of explanations and fairness guarantees. Interactive and user-centered approaches have been proposed to address this issue, but they introduce challenges related to usability and cognitive load [29].

In summary, future research should focus on developing unified multi-objective frameworks, improving data quality and bias handling, designing faithful and user-centered explanations, standardizing evaluation protocols, and ensuring scalable and privacy-preserving deployment. Addressing these challenges is essential for building recommender systems that are not only accurate, but also transparent, fair, and trustworthy.

8. Conclusion

In this survey, we presented a comprehensive and unified review of explainability and fairness in recommender systems, two critical dimensions for building trustworthy and responsible AI-driven platforms. While traditional recommender systems have primarily focused on predictive accuracy, recent advances highlight the importance of transparency and equitable outcomes as essential components of modern recommendation pipelines.

We began by introducing the fundamental concepts of recommender systems, explainability, and fairness, emphasizing their roles in enhancing user trust, system accountability, and ethical decision-making. We then proposed a multi-dimensional taxonomy that systematically organizes existing methods based on objectives, intervention stages, explanation types, and fairness perspectives. This taxonomy provides a structured view of the field and clarifies how different approaches address transparency and bias from complementary angles.

Furthermore, we reviewed state-of-the-art methods across three major categories: explainability techniques, fairness-aware approaches, and joint models that attempt to integrate both aspects. Our analysis shows that while significant progress has been made in each area, most existing methods still address explainability and fairness separately, with limited work on their joint optimization. In addition, we discussed evaluation metrics and highlighted the inherent trade-offs among accuracy, explainability, and fairness. Prior studies demonstrate that improving fairness or interpretability may affect predictive performance, underscoring the need for multi-objective optimization frameworks.

Based on this comprehensive analysis, we identified several open challenges, including data bias and incompleteness, lack of unified modeling frameworks, inconsistency in evaluation methodologies, and difficulties in real-world deployment. These challenges indicate that achieving explainable and fair recommender systems requires not only algorithmic innovation, but also advances in data-centric approaches, human-centered design, and standardized evaluation protocols.

Looking forward, we argue that the next generation of recommender systems should move beyond accuracy-centric optimization toward integrated, multi-objective frameworks that jointly consider transparency, fairness, and performance. Promising research directions include causal and debiasing-based models, interactive and user-adaptive explanation mechanisms, privacy-preserving fairness evaluation, and scalable solutions suitable for industrial deployment. Moreover, developing unified benchmarks and reproducible evaluation pipelines will be essential for advancing the field in a systematic and comparable manner.

In conclusion, explainability and fairness are not optional add-ons but fundamental requirements for the future of recommender systems. By bridging these dimensions within a unified framework, researchers and practitioners can design systems that are not only accurate, but also transparent, equitable, and trustworthy in real-world applications.

9. Conflicts of interest

The authors do hereby declare that there is no conflict of interests of other works regarding the publication of this paper. This article does not contain any studies with human participants or animals performed by any of the authors.

10. Data Availability

No new data were generated or analyzed in this study. All data discussed in this article are derived from previously published studies.

References

[1] Ai, J., Li, H., Su, Z., Zhao, F., 2025. An explainable recommendation algorithm based on content summarization and linear attention. Neurocomputing 630, 129692. doi:10.1016/j.neucom.2025.129692.

[2] Akkem, Y., Biswas, S.K., Varanasi, A., 2024. Streamlit-based enhancing crop recommendation systems with advanced explainable artificial intelligence for smart farming. Neural Computing and Applications 36, 20011–20025. doi:10.1007/s00521-024-10208-z.

[3] Almutairi, H., Ykhlef, M., 2026. Explainable reciprocal recommender system for affiliate–seller matching: A two-stage deep learning approach. Information doi:10.3390/info17010101.

[4] Alshammari, M., Nasraoui, O., Sanders, S., 2019. Mining semantic knowledge graphs to add explainability to black box recommender systems. IEEE Access 7, 110563–110579. doi:10.1109/ACCESS.2019.2934633.

[5] Amigó, E., Deldjoo, Y., Mizzaro, S., Bellogín, A., 2023. A unifying and general account of fairness measurement in recommender systems. Information Processing & Management 60, 103115. doi:10.1016/j.ipm.2022.103115.

[6] Arévalo, I., Salmeron, J.L., 2026. Model-agnostic post-hoc explainability for recommender systems. Expert Systems with Applications 298, 129608. doi:10.1016/j.eswa.2025.129608.

[7] Baghalzadeh Shishehgarkhaneh, M., Moehler, R., Fang, Y., Hijazi, A.A., Aboutorab, H., Dhamankar, A.J., 2026. Application of recommender systems in supply chain management: A state-of-the-art review. Intelligent Systems with Applications 30, 200661. doi:10.1016/j.iswa.2026.200661.

[8] Bojic, L., 2024. Ai alignment: Assessing the global impact of recommender systems. Futures 160, 103383. doi:10.1016/j.futures.2024.103383.

[9] de Brito Duarte, R., Correia, F., Arriaga, P., Paiva, A., 2023. Ai trust: Can explainable ai enhance warranted trust? Human Behavior and Emerging Technologies 2023, 4637678. doi:10.1155/2023/4637678.

[10] Burke, R., Ekstrand, M.D., Tintarev, N., Vassileva, J., 2021. Preface to the special issue on fair, accountable, and transparent recommender systems. User Modeling and User-Adapted Interaction 31, 371–375. doi:10.1007/s11257-021-09297-5.

[11] Caro-Martínez, M., Jiménez-Díaz, G., Recio-García, J.A., 2021. Conceptual modeling of explainable recommender systems: an ontological formalization to guide their design and development. Journal of Artificial Intelligence Research 71, 557–589. doi:10.1613/jair.1.12789.

[12] Caro-Martínez, M., Jiménez-Díaz, G., Recio-Garcia, J.A., 2023. A graph-based approach for minimising the knowledge requirement of explainable recommender systems. Knowledge and Information Systems doi:10.1007/s10115-023-01903-9.

[13] Charmchi, M.R.H., Cami, B.R., 2021. Paths-oriented test data generation using genetic algorithm, in: 2021 12th International Conference on Information and Knowledge Technology (IKT), IEEE. pp. 157–162. doi:10.1109/IKT54664.2021.9685262.

[14] Charmchi, M.R.H., Hassanpour, H., Cami, B.R., 2022. Considering user interests to provid an event base news stream framework, in: 2022 8th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), IEEE. pp. 1–5. doi:10.1109/ICSPIS56952.2022.10043990.

[15] Charmchi, M.R.H., Hassanpour, H., Rahimpour, B., 2024. Leveraging bot-connected user accounts for enhanced twitter (x) advertising outcomes. Journal of Artificial Intelligence, Applications and Innovations 1, 42–54.

[16] Chen, Q., Deng, J., Cheng, D., Li, J., Liu, L., 2026. Multi-view debiasing representation learning for recommender systems. Information Processing & Management , 104429doi:10.1016/j.ipm.2025.104429.

[17] Coba, L., Confalonieri, R., Zanker, M., 2022. Recoxplainer: A library for development and offline evaluation of explainable recommender systems, in: Proceedings of the 16th ACM Conference on Recommender Systems (RecSys), ACM. pp. 641–646. doi:10.1109/MCI.2021.3129958.

[18] De Campos, L.M., Fernández-Luna, J.M., Huete, J.F., 2024. An explainable content-based approach for recommender systems: a case study in journal recommendation for paper submission: Lm de campos et al. User Modeling and User-Adapted Interaction 34, 1431–1465. doi:10.1007/s11257-024-09400-6.

[19] De Filippo, A., Spillo, G., Boratto, L., Milano, M., Musto, C., Semeraro, G., 2026a. Recommender systems and sustainability: a dual perspective. Computer Science Review 60, 100912. doi:10.1016/j.cosrev.2026.100912.

[20] De Filippo, A., Spillo, G., Boratto, L., Milano, M., Musto, C., Semeraro, G., 2026b. Recommender systems and sustainability: a dual perspective. Computer Science Review 60, 100912. doi:10.1016/j.cosrev.2026.100912.

[21] Deldjoo, Y., Bellogin, A., Di Noia, T., 2021. Explaining recommender systems fairness and accuracy through the lens of data characteristics. Information processing & management 58, 102662. doi:10.1016/j.ipm.2021.102662.

[22] Deldjoo, Y., Jannach, D., Bellogin, A., Difonzo, A., Zanzonelli, D., 2024. Fairness in recommender systems: research landscape and future directions. User Modeling and User-Adapted Interaction 34, 59–108. doi:10.1007/s11257-023-09364-z.
[23] Duwe, G., Mercier, D., Kauth, V., Moench, K., Rajashekar, V., Junker, M., Dengel, A., Haferkamp, A., Hofner, T., 2025. Development of an artificial intelligence-generated, explainable treatment recommendation system for urothelial carcinoma and renal cell carcinoma to support multidisciplinary cancer conferences. European Journal of Cancer doi:10.1016/j.ejca.2025.115367.
[24] Gandhi, S., Balas, M., Popovic, M.M., Kertes, P.J., Muni, R.H., Kohly, R.P., 2026. Evaluation of algorithmic bias in large language models for retinal clinical recommendations. AJO International , 100233doi:10.1016/j.ajoint.2026.100233.
[25] Garapati, R., Chakraborty, M., 2025. Recommender systems in the digital age: a comprehensive review of methods, challenges, and applications. Knowledge and Information Systems 67, 6367–6411. doi:10.1007/s10115-025-02453-y.
[26] Gharahighehi, A., Vens, C., Pliakos, K., 2021. Fair multi-stakeholder news recommender system with hypergraph ranking. Information Processing & Management doi:10.1016/j.ipm.2021.102663.
[27] Gholami, S., Zarafshan, E., Sheikh, R., Sana, S.S., 2023. Using deep learning to enhance business intelligence in organizational management. Data science in Finance and Economics 3, 337. doi:10.3934/DSFE.2023020.
[28] Gómez, E., Contreras, D., Boratto, L., Salamó, M., 2025. Enhancing recommender systems with provider fairness through preference distribution awareness. International Journal of Information Management Data Insights doi:10.1016/j.jjimei.2024.100311.
[29] Guesmi, M., Chatti, M.A., Joarder, S., Ain, Q.U., Alatrash, R., Siepmann, C., Vahidi, T., 2024. Interactive explanation with varying level of details in an explainable scientific literature recommender system. International Journal of Human–Computer Interaction 40, 7248–7269. doi:10.1080/10447318.2023.2262797.
[30] Guttmann, M., Ge, M., 2024. Research agenda of ethical recommender systems based on explainable ai. Procedia Computer Science 238, 328–335. doi:10.1016/j.procs.2024.06.032.
[31] Gómez, E., Boratto, L., Salamó, M., 2022. Provider fairness across continents in collaborative recommender systems. Information Processing & Management 59, 102719. doi:10.1016/j.ipm.2021.102719.
[32] Hasan, E., Rahman, M., Ding, C., Huang, J.X., Raza, S., 2025. Based recommender systems: a survey of approaches, challenges and future perspectives. ACM Computing Surveys 58, 1–41. doi:10.1145/3742421.
[33] Hassanpour Charmchi, M.R., Hassanpour, H., Rahimpour, B., 2026. Integrating hybrid recommender systems and llms for twitter (x) campaigns: A framework for boosting engagement via bots and targeted content. International Journal of Engineering 39, 2911–2929. doi:10.5829/ije.2026.39.11b.19.
[34] Hou, Y., Yang, N., Wu, Y., Yu, P.S., 2019. Explainable recommendation with fusion of aspect information. World Wide Web 22, 221–240. doi:10.1007/s11280-018-0558-1.
[35] Huang, X., Chyczewski, T., Xia, Z., Kunz, R., Xiang, X., 2023a. Distilling experience into a physically interpretable recommender system for computational model selection. Scientific Reports doi:10.1038/s41598-023-27426-5.
[36] Huang, X., Chyczewski, T., Xia, Z., Kunz, R., Xiang, X., 2023b. Distilling experience into a physically interpretable recommender system for computational model selection. Scientific Reports doi:10.1038/s41598-023-27426-5.
[37] Huang, Z., Zhang, S., Cheng, D., Li, J., Liu, L., Lu, G., Zhang, G., 2026. Learning instrumental variable representation for debiasing in recommender systems. Neural Networks , 107977doi:10.1016/j.neunet.2025.107977.
[38] Jin, D., Wang, L., Zhang, H., Zheng, Y., Ding, W., Xia, F., Pan, S., 2023. A survey on fairness-aware recommender systems. Information Fusion 100, 101906. doi:10.1016/j.inffus.2023.101906.
[39] Khadivar, A., Khamseh, P.M., Basraei, R., Sana, S.S., 2026. Enhancing the price estimation in consulting services using empirically developed ann. International Journal of System Assurance Engineering and Management , 1–16doi:10.1007/s13198-025-03094-8.
[40] Koras, K., Kizling, E., Juraeva, D., Staub, E., Szczurek, E., 2021. Interpretable deep recommender system model for prediction of kinase inhibitor efficacy across cancer cell lines. Scientific Reports doi:10.1038/s41598-021-94564-z.
[41] Kumar, C., Alam, M., 2025. An in-depth analysis of recommender systems for integration of knowledge graphs utilising federated xai model. International Journal of Information Technology doi:10.1007/s41870-025-02491-3.
[42] Li, C., Ge, J., Cheng, K., Luo, B., Chang, V., 2022a. Statute recommendation: Re-ranking statutes by modeling case-statute relation with interpretable hand-crafted features. Information Sciences 606, 1023–1040. doi:10.1016/j.ins.2022.06.042.
[43] Li, J., Ren, Y., Sanderson, M., Deng, K., 2024a. Explaining recommendation fairness from a user/item perspective. ACM Transactions on Information Systems 43, 1–30. doi:10.1145/3698877.
[44] Li, W., Xu, Y., Li, Y., Huang, Y., 2025a. Display content, display methods, and evaluation methods of the hci in explainable recommender systems: A survey. Big Data Mining and Analytics 9, 198–228. doi:10.26599/BDMA.2025.9020049.
[45] Li, Y., Hedia, M.L., Ma, W., Lu, H., Zhang, M., Liu, Y., Ma, S., 2022b. Contextualized fairness for recommender systems in premium scenarios. Big Data Research 27. doi:10.1016/j.bdr.2021.100300.
[46] Li, Y., Liu, K., Satapathy, R., Wang, S., Cambria, E., 2024b. Recent developments in recommender systems: A survey [review article]. IEEE Computational Intelligence Magazine 19, 78–95. doi:10.1109/MCI.2024.3363984.
[47] Li, Z., Zou, L., Mao, C., Li, C., 2025b. Efficient and explainable sequential recommendation with language model. Information Processing & Management doi:10.1016/j.ipm.2025.104122.
[48] Liang, Q., Zheng, X., Wang, Y., Zhu, M., 2021. O3ers: An explainable recommendation system with online learning, online recommendation, and online explanation. Information Sciences doi:10.1016/j.ins.2020.12.070.
[49] Liu, H., Lin, H., Fan, W., Ren, Y., Xu, B., Zhang, X., Wen, D., Zhao, N., Lin, Y., Yang, L., 2022a. Self-supervised learning for fair recommender systems. Applied Soft Computing 122, 109126. doi:10.1016/j.asoc.2022.109126.
[50] Liu, H., Lin, H., Fan, W., Ren, Y., Xu, B., Zhang, X., Wen, D., Zhao, N., Lin, Y., Yang, L., 2022b. Self-supervised learning for fair recommender systems. Applied Soft Computing doi:10.1016/j.asoc.2022.109126.
[51] Liu, W., Liu, B., Qin, J., Zhang, X., Huang, W., Wang, Y., 2025. Fairness identification of large language models in recommendation. Scientific Reports 15, 5516. doi:10.1038/s41598-025-89965-3.

[52] Liu, Y., Yang, B., Pei, H., Li, J., 2020a. Neural explainable recommender model based on attributes and reviews. Journal of Computer Science and Technology doi:10.1007/s11390-020-0155-8.
[53] Liu, Y.Y., Yang, B., Pei, H.B., Huang, J., 2020b. Neural explainable recommender model based on attributes and reviews. Journal of Computer Science and Technology 35, 1446–1460. doi:10.1007/s11390-020-0152-8.
[54] Markchom, T., Liang, H., Ferryman, J., 2023. Scalable and explainable visually-aware recommender systems. Knowledge-Based Systems doi:10.1016/j.knosys.2023.110258.
[55] Markchom, T., Liang, H., Ferryman, J., 2025. Review of explainable graph-based recommender systems. ACM Computing Surveys 58. doi:10.1145/3772273.
[56] Musto, C., Narducci, F., Lops, P., De Gemmis, M., Semeraro, G., 2019. Linked open data-based explanations for transparent recommender systems, Elsevier. pp. 93–107. doi:10.1016/j.ijhcs.2018.03.003.
[57] Naghiaei, M., Rahmani, H.A., Deldjoo, Y., 2022. Pycpfair: A framework for consumer and producer fairness in recommender systems. Software Impacts 13, 100382. doi:10.1016/j.simpa.2022.100382.
[58] Parraga, O., More, M.D., Oliveira, C.M., Gavenski, N.S., Kupssinskü, L.S., Medronha, A., Moura, L.V., Simões, G.S., Barros, R.C., 2025. Fairness in deep learning: A survey on vision and language research. ACM Computing Surveys 57, 1–40. doi:10.1145/3637549.
[59] Paz-Ruza, J., Alonso-Betanzos, A., Guijarro-Berdiñas, B., Cancela, B., Eiras-Franco, C., 2024a. Sustainable transparency on recommender systems: Bayesian ranking of images for explainability. Information Fusion 111, 102497. doi:10.1016/j.inffus.2024.102497.
[60] Paz-Ruza, J., Alonso-Betanzos, A., Guijarro-Berdiñas, B., Cancela, B., Eiras-Franco, C., 2024b. Sustainable transparency on recommender systems: Bayesian ranking of images for explainability. Information Fusion 111, 102497. doi:10.1016/j.inffus.2024.102497.
[61] Ravi, M., Negi, A., Bommi, N.S., Rouf, N., 2025. Evolution of ai-driven decision making with decision support systems, expert systems, recommender systems, and xai. IETE Technical Review 42, 428–465. doi:10.1080/02564602.2025.2512086.
[62] Romero Meza, L., D'Urso, G., 2024. User's dilemma: A qualitative study on the influence of netflix recommender systems on choice overload. Psychological Studies 69, 349–367. doi:10.1007/s12646-024-00755-0.
[63] Roy, D., Dutta, M., 2022. A systematic review and research perspective on recommender systems. Journal of Big Data 9, 59. doi:10.1186/s40537-022-00592-5.
[64] Salimparsa, M., Sedig, K., Lizotte, D.J., Abdullah, S.S., Chalabianloo, N., Muanda, F.T., 2025. Explainable ai for clinical decision support systems: literature review, key gaps, and research synthesis, in: Informatics, MDPI. p. 119. doi:10.3390/informatics12040119.
[65] Salmerón, J.L., Arévalo, I., 2026. Spectral biclustering-driven scalability for post-hoc explainability in recommender systems. Knowledge-Based Systems , 115904doi:10.1016/j.knosys.2026.115904.
[66] Sardianos, C., Varlamis, I., Chronis, C., Dimitrakopoulos, G., Alsalemi, A., Himeur, Y., Bensaali, F., Amira, A., 2021. The emergence of explainability of intelligent systems: Delivering explainable and personalized recommendations for energy efficiency. International Journal of Intelligent Systems 36, 656–680. doi:10.1002/int.22314.
[67] Shafiloo, R., Stratigi, M., Peltonen, J., Olsson, T., Stefanidis, K., 2026a. The many facets of fairness in recommender systems: Consumers, providers and items. Information Systems doi:10.1016/j.is.2025.102643.
[68] Shafiloo, R., Stratigi, M., Peltonen, J., Stefanidis, K., 2026b. Multisided fairness under limited item availability in recommender systems. Information Sciences doi:10.1016/j.ins.2025.122926.
[69] Shastri, S., Kumar, S., Mansotra, V., et al., 2025. Advancing crop recommendation system with supervised machine learning and explainable artificial intelligence. Scientific Reports 15, 25498. doi:10.1038/s41598-025-07003-8.
[70] Shishehgarkhaneh, M.B., Moehler, R., Fang, Y., Hijazi, A.A., Aboutorab, H., Dhamankar, A.J., 2026. Application of recommender systems in supply chain management: A state-of-the-art review. Intelligent Systems with Applications , 200661doi:10.1016/j.iswa.2026.200661.
[71] Sivamayilvelan, K., Rajasekar, E., Vairavasundaram, S., Balachandran, S., Suresh, V., 2025. Building explainable artificial intelligence for reinforcement learning based debt collection recommender system using large language models. Engineering Applications of Artificial Intelligence doi:10.1016/j.engappai.2025.111622.
[72] Soremekun, E., Papadakis, M., Cordy, M., Le Traon, Y., 2025. Software fairness: An analysis and survey. ACM Computing Surveys 58, 1–38. doi:10.1145/3762170.
[73] Tang, J., Zhang, J., Tian, Z., Feng, X., Wang, L., Chen, X., 2025. Explainable recommendation with simulated human feedback. ACM Transactions on Information Systems doi:10.1145/3758091.
[74] Tejaswi, S., Sastry, V.N., Bhavani, S.D., 2026. Recommender system for secure mobile application based on permission pairs using explainable artificial intelligence. Expert Systems with Applications doi:10.1016/j.eswa.2025.130101.
[75] Vassøy, B., Langseth, H., 2024. Consumer-side fairness in recommender systems: a systematic survey of methods and evaluation. Artificial Intelligence Review 57, 101. doi:10.1007/s10462-023-10663-5.
[76] Vlachos, M., Dunner, C., Heckel, R., Vassiliadis, V., Parnell, T., Atasu, K., 2019. Addressing interpretability and cold-start in matrix factorization for recommender systems. IEEE Transactions on Knowledge and Data Engineering doi:10.1109/TKDE.2018.2829521.
[77] Vultureanu-Albişi, A., Murareţu, I., Bădică, C., 2025. Explainable recommender systems through reinforcement learning and knowledge distillation on knowledge graphs. Information 16, 282. doi:10.3390/info16040282.
[78] Wang, L., Xu, M., Zhang, Q., Shi, Y., Wu, Q., 2025a. Causal disentanglement for regulating social influence bias in social recommendation. Neurocomputing doi:10.1016/j.neucom.2024.129133.
[79] Wang, R., Tian, S., Wang, Z., Wang, Y., Zhang, D., Jiang, J., Yao, Y., Chen, H., Fang, H., Cao, M., 2025b. A clinical treatment recommender system optimizing adjuvant chemoradiotherapy benefit in chinese women with breast cancer using interpretable causal bayesian networks. The Lancet Regional Health - Western Pacific doi:10.1016/j.lanwpc.2024.101343.
[80] Wang, S., Zhang, X., Wang, Y., Ricci, F., 2024. Trustworthy recommender systems. ACM Transactions on Intelligent Systems and Technology 15, 1–20. doi:10.1145/3627826.
[81] Xu, C., Zhang, Y., Chen, H., Dong, L., Wang, W., 2023. A fairness-aware graph contrastive learning recommender framework for social tagging systems. Information Sciences 640, 119064. doi:10.1016/j.ins.2023.119064.

[82] Yaganteeswarudu, A., Biswas, S.K., Aruna, V., Tripathi, D., 2025. Enhancing transparency in smart farming: Local explanations for crop recommendations using lime. Procedia Computer Science doi:10.1016/j.procs.2025.04.450.
[83] Yang, N., Jo, J., Jeon, M., Kim, W., Kang, J., 2022. Semantic and explainable research-related recommendation system based on semi-supervised methodology using bert and lda models. Expert Systems with Applications doi:10.1016/j.eswa.2021.116209.
[84] Zaizi, F.E., Qassimi, S., Rakrak, S., 2025. Multi-objective optimization for personalized and fairness recommender systems: A hybrid approach. SN Computer Science 6, 713. doi:10.1007/s42979-025-04230-8.
[85] Zaizi, F.E., Qassimi, S., Rakrak, S., 2026. Multi-objective immune computing for recommender systems: A topsis-based optimization approach. The Review of Socionetwork Strategies doi:10.1007/s12626-026-00215-y.
[86] Zanjani, M.D., Aghdam, M.H., 2024. The explainable structure of deep neural network for recommendation systems. Future Generation Computer Systems 159, 459–473. doi:10.1016/j.future.2024.05.036.
[87] Zeng, J., Tao, H., Wen, J., Gao, M., 2025. Explainable next poi recommendation based on spatial–temporal disentanglement representation and pseudo profile generation. Knowledge-Based Systems 309, 112784. doi:10.1016/j.knosys.2024.112784.
[88] Zeyauddin, M., Abidin, S., Bokhari, M.U., Yadav, G., 2025. Toward transparent diagnosis of fatty liver disease: Explainable ai-driven recommender systems using shap and lime. International Journal of Information Technology doi:10.1007/s41870-025-02716-5.
[89] Zhang, Y., Jiang, J., Sun, J., Chan, H., Qiu, A., Liu, C., 2025. Hkgat: Heterogeneous knowledge graph attention network for explainable recommendation system. Applied Intelligence doi:10.1007/s10489-025-06446-w.
[90] Zhao, Z., Fan, W., Li, J., Liu, Y., Mei, X., Wang, Y., Wen, Z., Wang, F., Zhao, X., Tang, J., et al., 2024. Recommender systems in the era of large language models (llms). IEEE Transactions on Knowledge and Data Engineering 36, 6889–6907. doi:10.1109/TKDE.2024.3392335.
[91] Zhu, J., He, Y., Zhao, G., Bu, X., Qian, X., 2023. Joint reason generation and rating prediction for explainable recommendation. IEEE Transactions on Knowledge and Data Engineering doi:10.1109/TKDE.2022.3146178.