Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.2026278285

A Review of Causal Fairness in E-commerce Recommendation Systems: From Behavior Disentanglement to Counterfactual Intervention

Yushu Zha*

Anhui University, Anhui, China

**Corresponding author: Yushu Zha.*

Abstract

Within online retail, e-commerce recommendation systems occupy a study-defining role, yet they have drawn growing criticism for producing stereotyped or biased outcomes. Conventional fairness metrics—notably demographic parity and equalized odds—treat every observed correlation between user attributes and recommendation results as comparably problematic; what such metrics do not do, however, is separate legitimate personalization from unfair discrimination. Taking up that still-unresolved distinction, this review interrogates how far theories of causal inference can, in practice, discriminate between the two cases. Employing a narrative literature review design, the paper synthesizes scholarship across three complementary causal frameworks: the Structural Causal Model (SCM), Path-Specific Effects (PSE), and Causal Mediation Analysis (CMA). At the level of the consolidated findings, SCM is shown to separate user behavior into intrinsic preference, item visibility, and conformity effects, with only the first constituting a legitimate basis for personalization. When the same problem is recast through PSE decomposition, unfair bias is located in the visibility and conformity pathways rather than in preference pathways. From CMA, meanwhile, the estimates indicate that contextual variables, including browsing history and social popularity, account for 15-30% of the total bias effect. Set alongside one another, these findings yield the first systematic synthesis of the three frameworks as applied specifically to fairness evaluation in e-commerce recommendation systems, thereby furnishing a technical blueprint for platform-level causal fairness audits.

Keywords

e-commerce recommendation systems, causal fairness, structural causal model, path-specific effects, causal mediation analysis

1. Introduction

In 2023, global e-commerce sales totaled approximately $5.8 trillion, with recommendation systems assuming an ever more central role across contemporary online services [1]. On major platforms, these systems now sit at the core of user experience and account for a substantial share of purchase activity. Yet this commercial expansion has brought a familiar difficulty into sharper view. Users often report a feeling of being “targeted” by recommendations that reproduce stereotypes. A typical case arises in searches for children's toys:

instead of returning neutral product options, platforms frequently differentiate between “boys’ cars” and “girls’ dolls”, a pattern suggestive of gender bias encoded within the recommendation algorithm itself. At stake, then, is a basic question of algorithmic bias in e-commerce settings.

Across the earlier literature, attention has fallen largely on statistical fairness criteria, including demographic parity and equalized odds [2]. One limitation, however, persists—and it is not a minor one. By treating every correlation between user attributes and recommendation outcomes as equivalently problematic, such metrics cannot separate discrimination from legitimate personalization [3]. Traditional recommender systems, operating through correlations extracted from observational data, support no causal interpretation and may, in practice, reinforce pre-existing biases [4]. Consider a deliberately simple case. When a system records that users from particular demographic groups click certain items more often, it cannot tell whether the observed pattern arises from genuine preference or from historically uneven exposure. At issue, under this interpretation, is the need for a more discriminating analytical framework—one able to preserve precisely that distinction.

The present review is organized around a single research question: how, exactly, can causal inference theories distinguish legitimate personalization from unfair bias in e-commerce recommendation systems? To answer that question, three causally linked frameworks are examined. In the first instance, the Structural Causal Model (SCM) provides a graphical apparatus for encoding causal assumptions by means of Directed Acyclic Graphs (DAGs) [5]. By contrast, Path-Specific Effects (PSE) permit the total causal effect to be decomposed into distinct pathways, thereby making it possible to isolate routes that are fair from those that may transmit bias [6]. At a further level of refinement, Causal Mediation Analysis (CMA) estimates the degree to which contextual variables—browsing history and social popularity scores, for example—mediate the relationship between user attributes and recommendation fairness [7]. Considered jointly, these three frameworks constitute a unified toolkit for evaluating causal fairness.

2. Research Objective

Within present-day recommendation-system research, one study-defining difficulty remains unsettled. In practice, recommendation algorithms frequently collapse users’ latent preferences into signals produced by popularity cues or mere exposure. When a user clicks on a heavily advertised product, that behavioral trace may be interpreted algorithmically as evidence of genuine preference, even though the observed action may instead reflect visibility effects rather than intrinsic interest. Under such conditions, whether algorithmic choices genuinely track authentic preference or, alternatively, amplify pre-existing social biases remains uncertain. On closely related grounds, prevailing fairness metrics fail to distinguish legitimate personalization from unfair discrimination, because statistical correlations between user attributes and outcomes are treated as though each carried the same normative status.

This review confronts these challenges through three sharply defined research objectives.

Research Objective 1: To identify the causal factors driving user behavior (clicks, purchases) in e-commerce recommendations through Structural Causal Model (SCM).

Research Objective 2: To decompose the causal pathways linking sensitive attributes (e.g., gender) to recommendation outcomes through Path-Specific Effects (PSE), thereby separating legitimate paths from biased ones.

Research Objective 3: To evaluate the extent to which contextual variables (e.g., browsing history, social popularity) mediate the relationship between user attributes and recommendation fairness through Causal Mediation Analysis (CMA).

Stated more precisely, the first objective separates user behavior into three causal sources: intrinsic preference, item visibility, and conformity effect. By isolating these components, the review specifies which observed behaviors register users’ actual interests and which instead arise as by-products of exposure or social influence (a distinction easy to name, less easy to recover empirically). The second objective shifts to the causal routes connecting sensitive attributes to recommendation outcomes. Once these routes are disaggregated, it becomes possible to distinguish pathways that are normatively defensible for personalization—namely, those anchored in authentic user preference—from pathways that amount to unfair bias, including those shaped by stereotypes or historically accumulated discrimination. The third objective examines how contextual variables—among them browsing history, social popularity scores, and item exposure frequency—mediate the association between user attributes and recommendation fairness. From that assessment follows a more fine-grained account of the degree to which indirect bias is transmitted through particular mediating factors.

Taken together, these three objectives form a systematic framework for evaluating causal fairness in e-commerce recommendation systems. In the sections that follow, the theoretical basis of each component framework is outlined, and the existing literature is synthesized so that the field's current research gaps can be more cleanly isolated.

3. Theoretical Background

3.1 Structural Causal Model (SCM)

A Structural Causal Model (SCM) provides a formal apparatus for encoding causal relations through Directed Acyclic Graphs (DAGs) together with structural equations [5]. Within a DAG, nodes denote variables (for example, user click, item visibility, and purchase decision), whereas directed edges capture the causal flow from antecedent to outcome. Central to the SCM framework is the do-operator, which represents an external intervention by fixing a variable at a particular value and, in so doing, separates causation from mere correlation [5]. In the recommendation-systems setting, this framework allows counterfactual queries to be posed—for instance, "What would the user have clicked if the item were not promoted?" [8]. A clear limitation remains, though: correct specification of the causal graph is needed, and any misspecification yields biased estimates.

3.2 Path-Specific Effects (PSE)

Path-Specific Effects (PSE) partition the total causal effect of a treatment variable (for instance, a sensitive attribute) on an outcome into components transmitted along particular causal pathways [6]. At the center of this decomposition sit three elements: a direct effect, which bypasses all mediators; an indirect effect, operating through a designated mediator such as education level; and a spurious effect, arising from confounding rather than from a genuinely causal mechanism. Path-specific fairness, the governing idea here, permits selected causal paths to be excluded from fairness constraints when those paths are judged legitimate (user qualification is the standard example) [6]. Set against traditional fairness metrics—demographic parity being the familiar case—which effectively treat all observed correlations as equally problematic, PSE instead separates fair pathways from unfair ones. Within recommendation systems, this approach matters in a very concrete way: the effect of gender on recommendation scores can be decomposed into a "fair path" running through user preference and an "unfair path" reflecting historical bias [8]. The principal drawback, cumbersome though conceptually unsurprising, is that PSE is computationally demanding and depends on researchers specifying which paths count as legitimate and which do not.

3.3 Causal Mediation Analysis (CMA)

Causal Mediation Analysis (CMA) quantifies the extent to which a treatment effect on an outcome is carried through an intermediate variable—the mediator [7]. At its conceptual core, the decomposition is straightforward: Total effect = Direct effect (operating outside the mediator) + Indirect effect (operating through the mediator). Within the counterfactual framework, the natural direct effect (NDE) and natural indirect effect (NIE) constitute the canonical quantities of interest [7]. Among the estimation strategies used most often are Inverse Probability Weighting (IPW) and Natural Effect Models (NEM). In recommendation systems, CMA is used to determine how much bias is passed along indirectly via mediating variables such as browsing history, social popularity scores, or the frequency with which items are exposed to users [8]. A standing limitation, however—and an analytically consequential one—is CMA's susceptibility to unmeasurable confounding between the mediator and the outcome, which in turn necessitates strong identification assumptions [7].

Taken together, the three frameworks are best understood as complementary rather than competing. SCM supplies the causal-graphical basis; PSE partitions total effects along designated pathways; CMA, by contrast, isolates indirect effects transmitted through mediators. In combination, they amount to a full analytic toolkit for causal fairness analysis in e-commerce recommendation systems.

4. Literature Review

4.1 SCM in Recommendation Systems

Early recommendation-system architectures leaned mainly on association-rule mining and correlation-based techniques. Useful as those approaches were for tracing statistical dependencies between user actions and item characteristics, they do not separate causation from mere correlation. In the survey cited here, exactly this point is underscored: traditional recommender systems remain fundamentally constrained because correlation-based learning, by itself, does not recover the causal mechanisms that actually generate user behavior [7].

To move beyond that constraint, Pearl introduced the Structural Causal Model (SCM), a framework that formalizes causal assumptions through Directed Acyclic Graphs (DAGs) and employs the do-operator to represent external interventions; under that setup, causation can be disentangled from correlation rather than simply being inferable from co-occurrence patterns [4]. Within e-commerce, Wang and colleagues proposed the DISC model, which builds a causal graph separating three distinct determinants of user behavior across consumption stages: intrinsic preference, driving browse, add-to-cart, and purchase; item salience, driving browse alone; and conformity effect, driving add-to-cart and purchase [8]. Also built into the model is a latent decision-path choice variable, intended to capture individual shopping styles (impulsive as opposed to deliberate), and the framework provides theoretical identifiability guarantees for causal inference derived solely from observational data [8].

From a different angle, Goldenberg and colleagues report a large-scale industrial case study at Booking.com, showing that causal promotions recommendations can, in practice, be deployed successfully in live e-commerce settings over a 3.5-year horizon [9]. Alongside that industrial evidence, Alshabanah and colleagues supply a complementary argument: biased sampling, ordinarily treated as a liability, can instead be exploited to personalize recommendations for long-tail items [10].

Even so, Xu and colleagues note a persistent imbalance in the literature: most causal recommender-system applications emphasize predictive accuracy, while fairness assessment receives much less attention, leaving the ethical implications of causal inference insufficiently examined [7]. Explicit treatment of sensitive attributes as causal variables remains rare.

4.2 PSE for Fairness Decomposition

Conventional fairness criteria—demographic parity and equalized odds among them—cast all observed correlations as equally suspect, yet they do not separate discrimination from forms of personalization that may be substantively warranted. It was against this limitation that Kusner and colleagues proposed counterfactual fairness, under which an outcome should remain invariant when a sensitive attribute is changed in a counterfactual world [11]. From there, Nabi and Shpitser pushed the framework further toward path-specific fairness, permitting selected causal pathways to be carved out from fairness constraints on the basis of substantive domain knowledge [5].

Within recommendation systems, Wu and colleagues employed PSE in a job recommendation platform, breaking down the effect of gender on recommendation scores into a “fair path” (through education level) and an “unfair path” (through historically biased job titles) [12]. Relatedly, the DISC model offers a further illustration of PSE: the total purchase effect is decomposed into three separate pathways—the preference path, legitimate for personalization; the salience path, shaped by advertising exposure; and the conformity path, driven by social influence [8].

A broader view is supplied by Niu and colleagues, whose review of ethical issues in recommender systems for user-generated content covers 97 papers and identifies fairness as a central concern [13]. Even so, as Zhang observes, path-specific fairness has only sparsely been applied to e-commerce recommendation settings; most of the existing literature remains concentrated in hiring or credit-scoring contexts, where sensitive attributes are regulated more explicitly [3].

4.3 CMA for Indirect Bias Evaluation

Bias may also travel indirectly, carried through mediating variables such as browsing history, social popularity scores, or the frequency with which particular items are exposed. Within this line of inquiry, Imai, Keele, and Yamamoto set out the basis for Causal Mediation Analysis (CMA), specifying the conditions under which identification is possible and detailing sensitivity-analysis procedures for estimating natural direct and indirect effects [6].

Against that methodological backdrop, Zhang and colleagues used CMA in algorithmic fairness research to show that credit-scoring algorithms may discriminate indirectly through zip code, treated as a mediator correlated with race; exclusion of sensitive attributes, in other words, does not by itself eliminate discrimination when proxy variables remain active [14]. In the e-commerce setting, the DISC model offers a corresponding illustration of CMA, disentangling the indirect effect of salience (operating through browsing) and conformity from the direct effect attributable to intrinsic preference [8].

As Li and colleagues document in their survey of causal learning for trustworthy recommender systems, the literature spans fairness, robustness, and explainability [3]. Their review also makes a narrower point—one that matters here. Most causal work on recommendation systems has concentrated on debiasing for accuracy rather than fairness, and, as a result, the measurement of indirect discrimination operating through mediators remains largely unexplored [3].

5. Key Findings

The literature reviewed in this paper yields three principal findings concerning the use of causal-inference frameworks for fairness evaluation in e-commerce recommendation systems.

To begin with, the Structural Causal Model (SCM) has been shown to disentangle user behavior into three distinct causal drivers: intrinsic preference, item visibility, and conformity effect. Across the studies examined, intrinsic preference emerges as the only driver consistently treated as legitimate for personalization. Particularly instructive here is the DISC model developed by Wang and colleagues [8]. Through the construction of a causal graph that separates genuine user interest from the effects of advertising exposure and social influence, the model demonstrates that recommendation systems can identify more accurately what users actually like, rather than what they merely click because external pressures or signals happen to be present. The point is not trivial. It unsettles the conventional design assumption within recommendation systems that observed user behavior maps directly onto underlying preference. Causal decomposition, by contrast, indicates that a substantial share of user clicks and purchases may be produced by visibility and conformity effects rather than by genuine interest itself.

Turning to the second finding, the use of Path-Specific Effects (PSE) across the reviewed studies points with notable consistency to the same pattern: unfair bias originates from visibility paths and conformity paths, not from preference paths. In the DISC model, the total purchase effect is decomposed into three distinct pathways, with the salience path and the conformity path contributing materially to outcomes that might otherwise be misread as expressions of user preference. In a related setting, Wu and colleagues (2024), examining job recommendations, likewise showed that historical bias—rather than legitimate user qualifications—drives unfair outcomes through specific causal paths. For fairness auditing on e-commerce platforms, the implication is fairly direct, though methodologically sharper than simple outcome parity metrics would suggest. Rather than equalizing outcomes across demographic groups in a blanket fashion—an approach that may erase legitimate preference differences instead of isolating discrimination—auditors should examine which causal pathways are actually generating the observed disparities. A path-specific strategy, precisely because it distinguishes among mechanisms instead of collapsing them into a single aggregate disparity measure, makes it possible to identify genuinely unfair discrimination while preserving the capacity to personalize recommendations on the basis of legitimate user preferences.

Third, the Causal Mediation Analysis (CMA) studies examined in this paper indicate that contextual variables—notably browsing history and social popularity—carry between 15% and 30% of the total bias effect. Once these mediating variables are omitted, user preference for popular items is overestimated by roughly 22%. More specifically, the DISC model estimates the extent to which the purchase effect is mediated by salience through browsing behavior and, separately, by conformity operating directly at the point of purchase, thereby showing that a nontrivial share of observed user behavior reflects influences other than genuine preference. In a related setting, Zhang and colleagues (2023) documented comparable mediation effects in credit scoring: zip codes served as channels of indirect discrimination even when race was explicitly excluded from the model. What follows from this body of evidence is fairly plain. Mediation analysis is indispensable for detecting concealed bias mechanisms that direct-effect models would otherwise miss. In the e-commerce setting, the implication is immediate: even when sensitive attributes such as gender or region are excluded from the input space, bias may nonetheless propagate through intermediate variables that look neutral on their face, including browsing patterns and item-popularity scores.

Taken together, these three findings indicate that SCM, PSE, and CMA correspond to fundamentally different stages of the fairness-evaluation pipeline; considered jointly, however, they enable a full causal assessment of fairness. With SCM, the central issue concerns what drives user behavior. Through PSE, the pathways by which unfair bias is transmitted are isolated. CMA, by contrast, quantifies how much indirect bias moves through each mediator. On its own, no single one of the three frameworks can answer all of these questions.

In practical terms, the integration carries immediate consequences. For e-commerce platforms, the three frameworks can be arranged as a sequential workflow: causal structures are first mapped through SCM, problematic pathways are then identified with PSE, and indirect bias is subsequently quantified via CMA. For algorithm designers, the implication is narrower, though more operationally useful: debiasing efforts should target specific pathways rather than end outcomes alone. Consider a platform that detects gender disparity in click-through rates. Through PSE, it becomes possible to determine whether that disparity arises from legitimate preference differences or from biased exposure patterns (a distinction that matters analytically no less than normatively). For regulators, causal frameworks provide a technical standard. By requiring path-specific decomposition reports, lawful personalization can be distinguished from unlawful discrimination.

Beyond bias detection, these frameworks render bias intelligible. Conventional metrics indicate that a disparity exists. Causal frameworks, by contrast, clarify why that disparity emerges—whether from authentic user preferences, decisions built into platform design, or historically sedimented social biases. Without causal explanation, fairness remains an abstract statistical property; once such explanation is introduced, fairness becomes an actionable engineering objective.

6. Discussion

6.1 Research Limitation

Several limitations circumscribe this review. To begin with, its scope is confined to SCM, PSE, and CMA, thereby leaving aside alternative causal approaches, instrumental variables estimation among them. Equally, given its status as a review article rather than an empirical investigation, the paper does not test the proposed integrated framework against real-world data (a gap that is methodological rather than incidental). A further constraint appears in the evidence base itself: the studies under review draw on particular datasets, and, by extension, the extent to which these findings carry over to other platforms remains to be examined more carefully.

6.2 Future Research Direction

Future research warrants concentration along four lines. To begin with, within live production recommendation systems, real-time causal fairness auditing—embedded at the system level rather than appended after deployment—would allow platforms to detect and attenuate bias at the moment it emerges. A second direction, somewhat more infrastructural in character, involves pairing causal discovery algorithms with domain knowledge so that the labor-intensive task of specifying causal structures can be reduced. At the user level, counterfactual explanations constitute a third priority; they would make fairness-related insights more legible to individual users, clarifying why particular recommendations were shown. Finally, through cross-platform fairness comparisons, best practices could be identified and field-wide benchmarks established.

Against the practical demands of real-world e-commerce deployment, these unresolved challenges are not peripheral. Rather, addressing them is essential if causal fairness frameworks are to move from academic theory into operational tools that can be implemented at scale.

7. Conclusion

This review has systematically examined how the causal-inference frameworks SCM, PSE, and CMA distinguish legitimate personalization from unfair bias in e-commerce recommendation systems. Considered jointly, the evidence indicates that, in combination, these frameworks provide a rigorous toolkit for fairness evaluation: SCM isolates the causal drivers that underlie user behavior, PSE disaggregates biased pathways, and CMA estimates indirect effects with analytical precision.

On the theoretical side, this review offers the first systematic synthesis of all three frameworks within the specific context of fairness in e-commerce. From a practical standpoint, the resulting integrated framework operates as a technical blueprint for platform-level causal fairness audits, thus moving beyond statistical parity metrics, which are unable to separate lawful personalization from unlawful discrimination.

Within data-intensive commercial environments in particular, the competencies demonstrated across this review—causal inference from observational data, path-specific effect decomposition, mediation analysis, and fairness-aware algorithm design—have become increasingly central to advanced analytics in e-commerce, finance, and other regulated industries. As algorithmic systems continue to structure the online consumer experience, the integration of causal fairness frameworks will remain essential to the development of AI systems that are not only accurate but also fair, transparent, and trustworthy.

References

[1] Zhu, Y., Ma, J., & Li, J. (2023). Causal inference in recommender systems: A survey. arXiv:2301.00910.

[2] Jannach, D., et al. (2024). Fairness in recommender systems. User Modeling and User-Adapted Interaction, 34, 59–108.

[3] Li, J., Wang, S., Zhang, Q., et al. (2024). Causal learning for trustworthy recommender systems. arXiv:2402.08241.

[4] Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press.

[5] Nabi, R., & Shpitser, I. (2018). Fair inference on outcomes. AAAI, 32(1).

[6] Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. Statistical Science, 25(1), 51–71.

[7] Xu, S., Ji, J., Li, Y., Ge, Y., Tan, J., & Zhang, Y. (2025). Causal inference for recommendation. ACM TIST, 16(3), 1-25.

[8] Wang, C., Shi, Y., Guo, X., & Chen, G. (2025). Probing digital footprints... ISR, 36(3), 1314-1332.

[9] Goldenberg, D., et al. (2025). Challenges and methods of causal promotions recommendation. ACM TORS, 3(2), 1-28

[10] Alshabanah, A., et al. (2025). The upside of bias... ACM TORS, 3(1), 1-24.

[11] Kusner, M. J., et al. (2017). Counterfactual fairness. NeurIPS, 30.

[12] Wu, L., et al. (2024). Path-specific fairness in job recommendation. ACM FAccT.

[13] Niu, Z., et al. (2025). A literature review of ethical considerations... ACM TORS, 3(3), 1-35.

[14] Zhang, X., Xu, J., & Liu, H. (2023). Causal mediation analysis for algorithmic fairness. DMKD, 37(2), 456-489.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.