Bias Mitigation for AI-Feedback Loops in Recommender Systems: A Systematic Literature Review and Taxonomy

Theodor Stoecker
Technical University of Munich
Munich, Germany
theo.stoecker@tum.de

Samed Bayer
Technical University of Munich &
Fraunhofer Gesellschaft
Munich, Germany
samed.bayer@tum.de

Ingo Weber
Technical University of Munich &
Fraunhofer Gesellschaft
Munich, Germany
ingo.weber@tum.de

ABSTRACT

Recommender systems continually retrain on user reactions to their own predictions, creating AI feedback loops that amplify biases and diminish fairness over time. Despite this well-known risk, most bias mitigation techniques are tested only on static splits, so their long-term fairness across multiple retraining rounds remains unclear. We therefore present a systematic literature review of bias mitigation methods that explicitly consider AI feedback loops and are validated in multi-round simulations or live A/B tests. Screening 347 papers yields 24 primary studies published between 2019–2025. Each study is coded on six dimensions: mitigation technique, biases addressed, dynamic testing set-up, evaluation focus, application domain, and ML task, organising them into a reusable taxonomy. The taxonomy offers industry practitioners a quick checklist for selecting robust methods and gives researchers a clear roadmap to the field's most urgent gaps. Examples include the shortage of shared simulators, varying evaluation metrics, and the fact that most studies report either fairness or performance; only six use both.

1 INTRODUCTION

Recommender systems (RS) rely on dynamic machine learning (ML) models to personalise suggested content at scale. The algorithms continuously learn from the repeated interactions based on their own prior predictions [31] or, more generally, outputs [2]. This may lead to self-reinforcing AI-feedback loops, which, over successive cycles, can further amplify the prior bias in the system [20, 25]. Such bias amplification harms the recommendation diversity, system fairness, long-term platform health, and user trust [29].

Most bias mitigation approaches for RS are evaluated on a single iteration of the training/validation/testing data splits, ignoring the feedback loop effect, therefore introducing evaluation bias [25, 3]. According to a related survey, 115 studies out of 127 are evaluated using offline testing without considering model updates [25]. Recent work in the literature, however, indicates a shift towards dynamic bias auditing. Simulations that replay many retraining rounds show that bias-mitigation approaches, which initially succeed, can fail to mitigate the bias in the long term [1]. Frameworks such as FADE [39] and FairAgent [16] update models while enforcing fairness constraints. Another study documents the AI-feedback loop amplification of source bias when RS learn from AI-generated content [41].

Although recent surveys classify individual bias categories and mitigation strategies, the field still lacks a systematic, empirical overview that links biases with dynamic mitigation strategies that remain effective under continual learning with feedback loops. Our work addresses this gap with twofold contributions:

- We conduct a systematic literature review on bias mitigation strategies within ML Model feedback loops, tested with retraining in simulation or live environments.
- Based on the literature, we propose a taxonomy that organises bias mitigation techniques in recommender systems by mitigation type, biases addressed, dynamic testing type, evaluation focus, application domain and ML model task.

In RS, feedback loops can lead to biased and unfair decisions that threaten the long-term health of platforms. Thus, our work has implications for both academic researchers and industry practitioners.

2 BACKGROUND AND RELATED WORK

This section first outlines biases in RS and then summarises feedback-loop types, bias mitigation approaches, and existing surveys.

2.1 Bias in Recommender Systems

Machine learning models create predictions based on statistical patterns learned through observed data [4, 2]. In RS, these models are updated based on new data collected in user interaction steps such as user feedback for suggested items [31]. This leads to a loop of prediction, interaction, and retraining, where each step influences the others.

Suresh and Guttag categorise bias into seven types: *historical*, *representation*, *measurement*, *aggregation*, *learning*, *evaluation*, and *deployment* [35]. Moreover, Chen et al. present another influential taxonomy of bias types in RS with the feedback loop and seven classes: *selection bias*, *exposure bias*, *conformity bias*, *position bias*, *inductive bias*, *popularity bias*, and *unfairness* [9].

Following the feedback loop framework of Pagan et al. [31] (see Section 2.2), we decide to focus on four of Suresh and Guttag's bias classifications for our taxonomy: *representation*, *historical*, *measurement*, and *evaluation*. These are the ones that are most relevant to feedback loops. Figure 1 illustrates how evaluating a model in a (simulated) live setting exposes the recurring nature of these biases and how they can be amplified.

2.2 Biases in AI-Feedback Loops

Pagan et al. identified five types of feedback loops, characterised by their position within the ML system and the component affected: *Sampling feedback loop*, *Individual feedback loop*, *ML Model feedback loop*, *Feature feedback loop*, and *Outcome Feature Loop* [31].

ML Model feedback loops arise when a system retrains (or evaluates) itself on the very instances it has already considered as only belonging to a specific class, such as the relevance for a user [31]. Figure 1 shows how RS exemplify this concept: only recommended

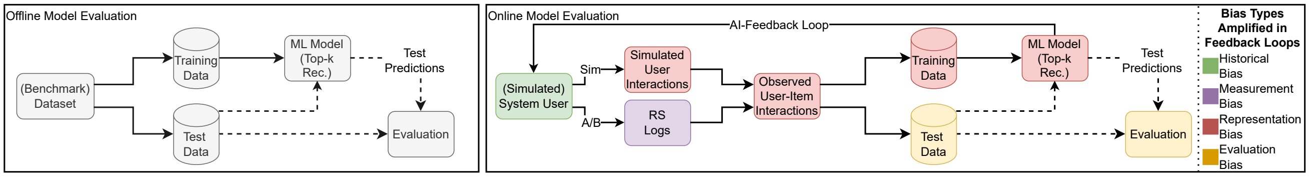


Figure 1: Offline evaluation detects bias in static data, whereas online (or simulated) evaluation reveals how feedback loops amplify four bias types—historical (green), measurement (purple), representation (red), and evaluation (yellow)—giving a more realistic view of model dynamics over time.

items receive user feedback. Adding this feedback to the training data amplifies representation bias, whereas adding it to the test set reinforces evaluation bias. Additional loops can arise—for example, Individual feedback loops in which users modify their preferences in response to the system's suggestions [31].

For the rest of this work, we differentiate the types of feedback loops based on their classification. We focus primarily on ML Model feedback loops in our literature review selection process, in order to ensure comparability of mitigation techniques. Unless explicitly naming a feedback loop type, feedback loops, including AI feedback loops, refer to ML Model feedback loops in later sections.

2.3 Bias Mitigation Techniques and Classifications

Related works often use a well established framework for bias mitigation classification based on their stage in the ML-pipeline: *Pre-*, *In-*, or *Post-Processing*. Pre-processing considers changes done before the model uses the data as input, in-processing describes an approach that changes some part of the prediction or learning process, and post-processing changes the model output [19, 7, 25].

Although the pipeline-based taxonomy is well established, its concrete subclasses vary across the literature (cf. [19, 7]). Both Hort et al. and Caton and Haas concentrate on classifiers and therefore overlook tasks such as causal inference and reinforcement learning that are commonly used in RS.

We include Caton and Haas' classification despite their limitations for RS due to their impact in the related field of ML fairness and classifiers. They identify 16 different classes within the three pipeline stages: *Adversarial Learning*, *Causal Methods*, *Relabelling and Perturbation*, *Resampling*, *Reweighing*, *Transformation*, and *Variable Blinding* for **Pre-Processing**; *Adversarial Learning*, *Bandits*, *Constraint Optimisation*, *Regularisation*, and *Reweighing* for **In-Processing**; and *Calibration*, *Constraint Optimisation*, *Thresholding*, and *Transformation* for **Post-Processing** [7]. Because many of those sub-classes either do not appear in or are not relevant to the feedback loop studies in RS, we introduce an additional set of sub-classes in the following sections.

2.4 Existing Surveys on Biases in AI Feedback Loops

We identify three prior surveys that also examine bias in RS, each from a different angle: a bias taxonomy for RS, a focused study on popularity bias, and a work on causal inference mitigation methods.

The closest related survey addresses seven different biases and the feedback loop effect in RS [9]. However, by developing distinct bias categories and adopting a single feedback loop concept, they offer a perspective that differs from the classification of feedback loops and biases outlined above. Another study examines a single bias—popularity bias—but surveys a broader range of mitigation methods, many of which are evaluated only in offline settings [25]. Consequently, its scope differs from ours in both the biases addressed and the evaluation approaches considered. Lastly, Li et al. consider causal inference mitigation techniques for RS [26]. While we too included studies on causal inference techniques, we also include other approaches, as explained in the prior section.

Our research is different from existing works by focusing specifically on the current state of bias mitigation strategies for ML Model feedback loops, evaluated in a dynamic environment, including model updates.

3 METHOD

We queried two scientific databases: *ACM Digital Library*, because it includes the most relevant conferences in the field, such as RecSys, WWW, or KDD; and *IEEE Xplore* for their relevance in technical fields as a complementary source. Other databases were excluded due to resource constraints. We added studies by two other methods: the inclusion of already identified ML Model feedback loop papers in the classification of Pagan et al. [31]; and an additional search on ArXiv, to also include the most up-to-date research on the topic, despite their lack of a full peer review.

Our inclusion criteria required i) the presence of an ML Model feedback loop (see Section 2.2), and ii) an applied approach of bias mitigation, tested in iii) a dynamic environment with updates of the model, such as in simulations or live-testing with more than one iteration of a loop, such as depicted in Figure 1. Only research papers from conferences, workshops, and journals were considered; extended abstracts, and posters were excluded.

The search strings shown in Table 1 were chosen to find relevant studies by finding ML related research on bias mitigation in feedback loops. As fields use different words to describe a similar phenomenon, we also included "echo chamber" to also consider related studies. This same search string, however, did not yield any papers for IEEE Xplore, so we queried a search without the explicit mention of ML or related words. For the ArXiv search, we conducted a simple search for feedback loops in ML in connection with bias mitigation published in the last six years. The most common reasons for exclusion were missing AI-feedback loops in screening stage 1 and a lack of an online evaluation in screening stage 2.

Table 1: Search details including the database, the last access time, the used search string, and the number of studies found.

Database	Last accessed	Search string	Studies found
IEEE Xplore	2025-06-13	("Abstract":"*feedback loop" OR "Abstract":"bias amplification" OR "Publication Title":"*feedback loop" OR "Publication Title":"bias amplification" OR "Author Keywords":"*feedback loop" OR "Author Keywords":"bias amplification") AND ("Full Text .AND. Metadata":mitigation OR "Full Text .AND. Metadata":guardrail*) AND ("Abstract":bias OR "Abstract":biases OR "Publication Title":bias OR "Publication Title":biases OR "Author Keywords":bias OR "Author Keywords":biases)	13
ACM DL	2025-05-23	query: {AllField:("Machine Learning" OR ML OR "Artificial intelligence" OR "AI" OR "Deep Learning") AND (Abstract:(bias*) OR Keyword:(bias*)) AND (Keyword:("*feedback loop*" OR "*bias amplification*" OR "*reinforc* feedback*" OR "echo chamber" OR "Recommender System") OR Abstract:("*feedback loop*" OR "*bias amplification*" OR "*reinforc* feedback*" OR "echo chamber" OR "Recommender System")) AND AllField:(Prevention OR mitigation OR "mitigation strategy" OR countermeasure OR control OR reduction)} "filter":	318
ArXiv	2025-08-25	order: -announced_date_first; size: 200; date_range: from 2019-06-01 to 2025-08-31; classification: Computer Science (cs); include_cross_list: True; terms: AND all="machine learning" OR ML OR "artificial intelligence" OR AI OR "deep learning"; AND all=bias; AND all=feedback loop; AND all=mitigation	11

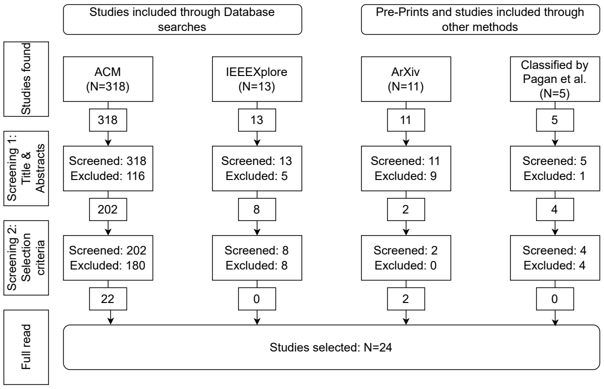


Figure 2: Flowchart of study selection. From 347 **records identified (ACM** 318**; IEEE Xplore** 13**; ArXiv** 11**; Pagan et al.** 5**), titles and abstracts were first screened for AI-feedback loops and bias mitigation; full texts were then assessed for applied mitigation strategies tested via simulation or A/B testing, yielding** 24 **studies.**

Our procedure to select relevant papers as outlined in Figure 2 was twofold. In a first screening, we searched for relevancy by examining title and abstract for a relation to AI-feedback loops, biases, and mitigation strategies. This led to 150 papers. To augment our selected texts and to account for personal bias, we utilised a large language model (LLM: *gpt-4o-mini*[1]) for texts from the ACM database, which identified 158 papers in this screening step (93 overlapping with our initial set). This yielded 216 non-duplicate papers for full-text assessment. The final selection decision remained human in the next step. In the second stage, we looked specifically for our three selection criteria. After full-text review, 24 papers met all criteria. A second evaluator independently checked this selection for consistency.

We constructed the taxonomy using Nickerson et al.'s framework through multiple conceptual-to-empirical iterations [30]. First, we began with established categories from prior surveys or related work (Section 2). We then iteratively mapped every candidate study onto the current taxonomy. When a study did not fit clearly to the current categories in the literature, we tailored category definitions or introduced a new sub-class (e.g., Add-On for architectural extensions). We ended the iterative process once all studies fit into the taxonomy, satisfying the objective and subjective stopping criteria outlined by Nickerson et al. [30]. Finally, we coded the full set of 24 studies with the final taxonomy in Table 2.

4 RESULTS

We start by presenting metadata about the studies, and continue with the criteria outlined in Section 3. The list of papers mapped to our criteria can be found in our table for reproducibility [2].

4.1 Study Metadata and Attributes

Figure 3 lists the publication venues, most of which are implementation oriented outlets. Across these venues, 22 of the 24 papers in our sample of ML Model feedback loop studies address RS. ACM's Conference on Recommender Systems (RecSys) is the most prominent one. Venues for fairness research, such as the ACM Conference on AI, Ethics, and Society (AIES), play only a smaller role. Just two of the included studies were published in a journal, which indicates the fast pace and conference-focused nature of the field.

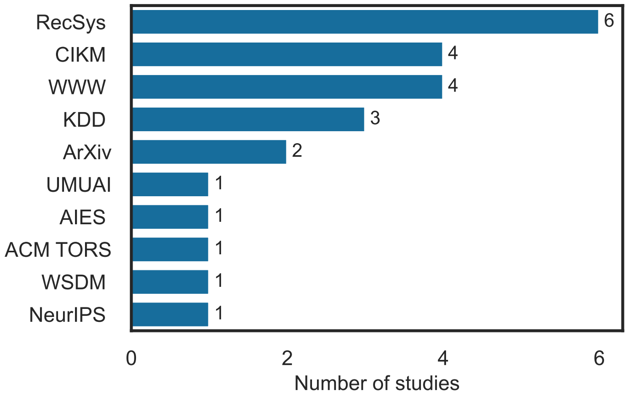


Figure 3: Number of included studies per publication venue ($N = 24$**). RecSys leads with six studies (25 %), followed by CIKM and WWW with four each (17 %).**

[1] See prompt used in Online Appendix

[2] See full list of papers in the Online Appendix

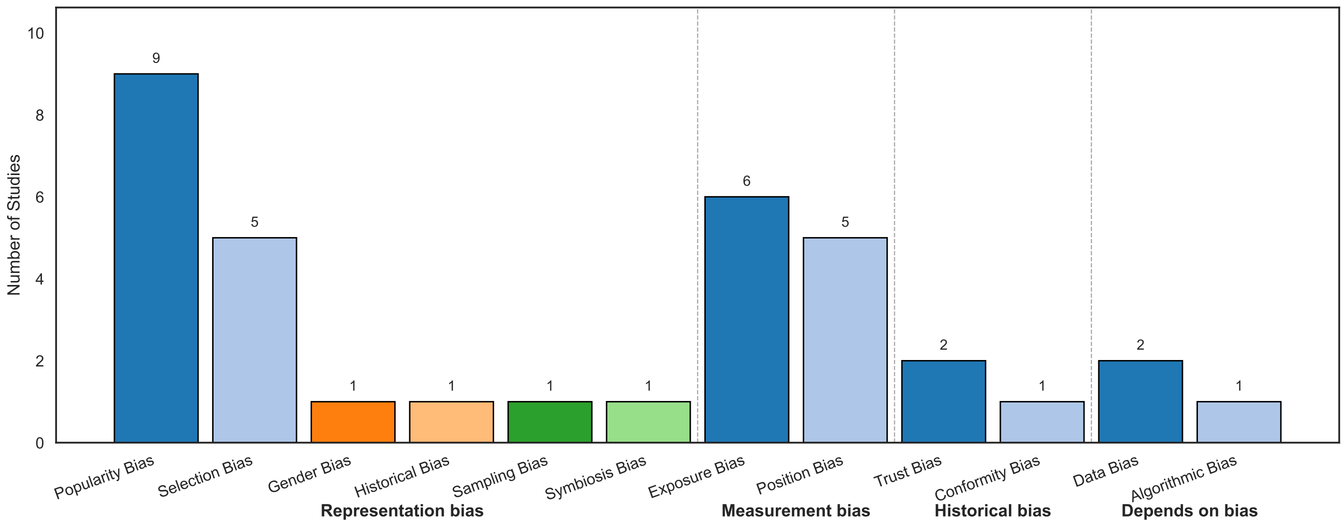


Figure 4: Biases mentioned and their mapping to Suresh and Guttag's framework. Data and algorithmic bias depend on the specific bias as these are too general terms to clearly classify. Please note that the representation category in a feedback loop can also lead to evaluation bias instead.

17 studies have at least one author affiliated with industry (71 %), with 13 of them having the majority of authors affiliated with industry (54 %). The extent of industry research, primarily from large platforms utilising RS, highlights the high demand in practice. Co-authored studies of industry and academia also indicate the close cooperation between them.

The publication years ranged from 2019 to 2025, with 2020 and 2022 having the most studies, five and six, respectively.

4.2 Domain of System Applications

While most of the papers considered general-purpose RS (9), the most popular domains are music (2), and movie/video recommendation (4). Domains of RS represented by only one paper fall into the "Other" category: E-Commerce, advertisements, social networks, app store, news, search recommendation, and A/B testing for recommendation systems. The two studies unrelated to RS focused on reinforcement learning, and on loan applications, healthcare, and policing. The variety of fields where RS are applied shows the applicability of mitigation approaches across different fields.

The different input data, ranging from text in news recommendation to music and video data, also highlights the need to find easily usable features for the RS. Examples of this are watch time, to extract the user interest of a video based on an easily measurable metric [27].

4.3 ML-Model tasks

The tasks for RS ranged from predicting specific metrics, candidate generation, to ranking the outputs. As mentioned in the prior section, two systems are completely independent of RS, their tasks focused on classification and optimisation of resource allocation between groups.

Top-k recommendation is the dominant task, addressed in 20 studies. Of these, seven concentrate on the ranking phase, and two treat the special case of $k = 1$—where user feedback differs, as preferences over a list of items are absent when only a single item is presented.

Another type of tasks has the aim of predicting features used in the RS (e.g., Click through rate (CTR) [15], or watch time prediction [27]). CTR and watch time are logged but unobserved at inference. By predicting these values from history, the model can treat them as features and improve the following top-k recommendation.

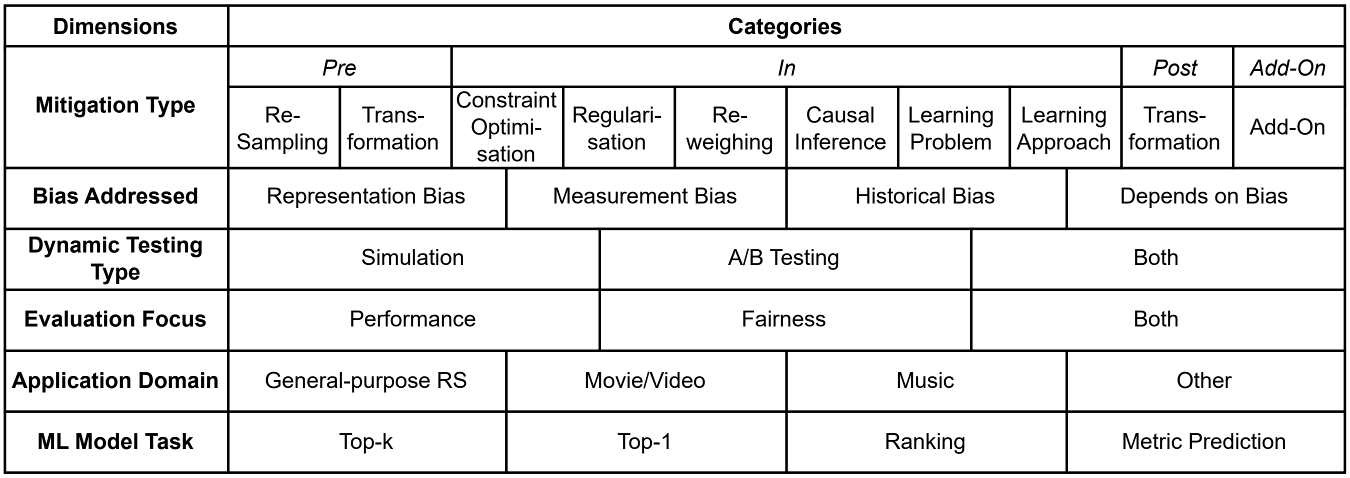
4.4 Biases addressed

Studies were motivated by different, sometimes multiple, biases. For each paper, we recorded every bias that was explicitly addressed. For papers that did not name a specific bias, we inferred the nearest matching bias from the problem statement when possible.

Apart from the five biases identified in the survey on biases in RS by Chen et al. [9], an additional seven were named explicitly in the studies. The mentioned biases and their mapping to the framework of Suresh and Guttag [35] are visualised in Figure 4. "Depends on bias" is used in this mapping whenever the mentioned bias is too broad to be classified. Furthermore, note that in the one case of historical bias, the description of the bias used in the study is different from its description in the classification framework [35].

4.5 Mitigation Types

Mitigation approaches relied on changes to the in-processing stage of a model in 17 out of 24 (71 %) cases, the pre-, and post-processing stage were changed in two and three studies respectively. Two studies changed the model or system architecture to mitigate biases, which is not covered in the established pre-/in-/post-processing. We categorised them as "Add-On" as a separate mitigation type in this study.



Dimensions	Categories									
Mitigation Type	*Pre*		*In*						*Post*	*Add-On*
	Re-Sampling	Trans-formation	Constraint Optimi-sation	Regulari-sation	Re-weighing	Causal Inference	Learning Problem	Learning Approach	Trans-formation	Add-On
Bias Addressed	Representation Bias			Measurement Bias		Historical Bias			Depends on Bias	
Dynamic Testing Type	Simulation			A/B Testing				Both		
Evaluation Focus	Performance			Fairness				Both		
Application Domain	General-purpose RS			Movie/Video		Music			Other	
ML Model Task	Top-k			Top-1		Ranking			Metric Prediction	

Figure 5: Taxonomy for recommender systems bias mitigation evaluated in dynamic environments based on six dimensions.

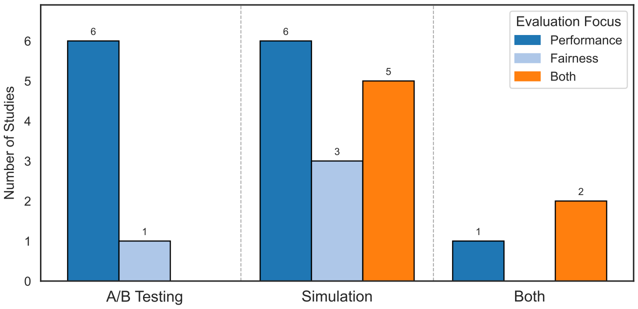


Figure 6: Metrics used to evaluate a mitigation type in simulated or A/B tested feedback loop environments. A/B tests mostly use performance type evaluations.

Of the presented subclasses identified by Caton and Haas [7], we identified six in the context of RS: *Resampling*, and *Transformation* in **Pre-Processing**; *Constraint Optimisation*, *Regularisation*, and *Reweighing* for **In-Processing**; and *Transformation* for **Post-Processing**.

To cover the remaining studies, we extend the taxonomy with four RS-specific classes: *Causal Inference-based* from the survey on popularity bias in RS [25], *learning approach* to capture a change to a zero-shot method, and *learning problem* for reinforcement-learning or reward or loss function changes **In-Processing**; and *Add-On*, which covers the addition of components rather than changing the established pipeline steps, such as adding a Monte Carlo simulation.

4.6 Evaluation and Metrics

We visualised evaluation types and used metrics in Figure 6. In the literature retraining is done by either using a simulation or live A/B tests. In simulations, a user-system-interaction is simulated, whereas in an A/B test, two groups of live users get recommendations from different versions of the RS (one with bias mitigation, the other one without, as a control group). 54 % of studies focused only on performance, 17 % on measuring the effect of their mitigation technique using a fairness metric, and the other 29 % investigated both performance and fairness. A/B Tests and industry works were most likely to be evaluated with performance metrics.

The most popular performance metrics were Click Through Rate (CTR) [17, 24, 15, 22, 21, 38], and normalised Distributed Cumulative Gain (nDCG) [38, 37, 28] or variations of those. Other metrics included some commonly used in machine learning, such as mean absolute/squared error (MAE/MSE) [21, 27], Area under the receiver operator curve (ROC-AUC) [14], or cumulative gain (for reinforcement learning) [36]. In one case, the performance metrics were adapted to specifically investigate the bias. Han et al. measured the overall performance, and newer items in particular, to see the effect on the cold-start problem [18].

Evaluations based on fairness measurements vary by use case. GINI Coefficient was the most popular choice [12, 13, 8]. Other than that, specific metrics are used, for instance: difference in granted loans between gender [33], or the position of first female artist for music recommendation [12].

5 TAXONOMY AND DISCUSSION

5.1 Taxonomy

As described in Section 3, we followed a well-established method [30] to derive a taxonomy – the result is visualised in Figure 5. Our six-dimensional taxonomy captures bias-mitigation approaches for RS that incorporate an ML Model feedback loop and are evaluated in dynamic environments. Table 2 classifies the 24 studies from our literature review using the proposed taxonomy.

First, the mitigation type, rooted in the established pipeline-stage classification and its sub-categories, indicates both where in the workflow the bias-reducing intervention takes place and what it specifically modifies. Next comes the bias type—the specific bias the authors highlight as their primary motivation. While related taxonomies covered additional classes of mitigation types [7, 19], they were unable to cover all identified types. We added four classes to classify those papers, as described in Section 4.5. The lack of the other classes in their studies might indicate potential gaps in

Table 2: Taxonomy mapping for each of the 24 papers across our seven core dimensions. Biases in brackets are mapped in cases without explicitly mentioned biases. The last two studies are the ones unrelated to recommender systems. Representation, measurement, and historical bias are abbreviated as Rep., Meas., and Hist., respectively.

Paper	Mitigation Type	Bias Addressed	Dynamic Testing Type	Evaluation Focus	Application Domain	ML Model Task
[18]	Transformation (Pre)	Rep., Meas., and Hist. Bias	Both	Both	Other (E-Commerce)	Ranking
[17]	Causal Inference-based	Rep. Bias	Sim	Performance	General-purpose RS	Top-k
[8]	Regularisation	Rep. Bias	A/B	Fairness	General-purpose RS	Top-k
[11]	Reweighing	Meas. and Rep. Bias	Sim	Both	General-purpose RS	Top-k
[14]	Learning Problem	Depends	Both	Performance	Other (Advertisement)	Top-k
[6]	Learning Problem	Meas. and Rep. Bias	Sim	Both	General-purpose RS	Top-k
[24]	Transformation (Post)	Rep. Bias	A/B	Performance	Movie/Video	Top-k
[13]	Transformation (Post)	Rep. Bias	Sim	Both	Music	Top-k
[12]	Transformation (Post)	Rep., and Meas. Bias	Sim	Fairness	Music	Ranking
[1]	Constraint Optimisation	Rep., Meas., and Hist. Bias	Sim	Fairness	Other (Social Network)	Top-k
[34]	Learning Problem	Meas. Bias	A/B	Performance	General-purpose RS	Ranking
[28]	Reweighing	(Meas. Bias)	Sim	Performance	General-purpose RS	Ranking
[37]	Reweighing	Meas., Hist., and Rep. Bias	Sim	Performance	General-purpose RS	Top-k
[15]	Add-On	Meas. Bias	A/B	Performance	Other (App Store)	Metric prediction
[21]	Learning Problem	Rep. Bias	Sim	Performance	General-purpose RS	Top-1
[22]	Learning Problem	Rep. Bias	Sim	Performance	General-purpose RS	Top-1
[5]	Resampling	Rep. Bias	Sim	Fairness	Other (A/B Testing for RS)	Ranking
[10]	Reweighing	Depends	Both	Both	Movie/Video	Top-k
[27]	Causal Inference-based	Rep. Bias	A/B	Performance	Movie/Video	Metric prediction
[32]	Reweighing	Meas., and Hist. Bias	A/B	Performance	Other (News)	Ranking
[38]	Learning Approach	(Rep. Bias)	A/B	Performance	Other (Search)	Top-k
[23]	Regularisation	Meas. Bias	Sim	Both	Movie/Video	Ranking
[36]	Learning Problem	Rep. Bias	Sim	Performance	Non Recommender System	Other (Optimisation)
[33]	Add-On	Depends	Sim	Both	Non Recommender System	Other (Classification)

literature, but could also be due to the number of selected studies fitting our selection criteria, as well as the difficulty to implement specific types, such as adversarial learning approaches, for RS. A third dimension tracks reiteration or dynamic testing, revealing how the model tries to show the effectiveness over a number of feedback loop iterations. Fourth, the evaluation focus tells us which metric is used to judge whether the performance of the changed model or system actually improves. Finally, we distinguish between the domain, the real-world application in which the work operates, and the ML model task, that is, what the model ultimately predicts in RS.

This taxonomy allows the classification of new bias mitigation approaches while staying compatible with established bias and feedback loop frameworks outlined in earlier sections.

5.2 Discussion

When considering biases in RS, it is important to note that feedback loops amplifying the popularity of items can also be beneficial—a high-quality item might be suggested to a larger number of users. The challenge here lies in identifying the origin of popularity: quality or system-introduced [40].

While ML Model feedback loops primarily consider representation and evaluation bias, some of the mentioned biases are classified as measurement or historical bias as well. This is because of three main reasons: Firstly, we confirm the finding of other studies that the usage of bias terminology is unclear within the literature [25, 31, 9]. In addition to this, the existence of different feedback loop types within one system, and the mention of multiple biases within a single study often obscures which specific biases are actually being examined. Lastly, for the bias mapping we rely on the used bias terminology, not on a single study basis.

6 CONCLUSION, LIMITATION AND FUTURE WORK

In summary, bias mitigation in feedback looped recommender systems is a growing field, dominated by in-processing mitigation techniques. We conducted a systematic literature review yielding 24 recent studies and introduced a taxonomy that helps categorise and compare bias mitigation techniques within feedback loops across six dimensions.

Our work is limited by the choice of mitigation approaches considered. Specifically, the databases used and the need for dynamic testing in the evaluation of the mitigation approach. Another limitation concerns the databases we used; in future work, further databases could be included to get a more holistic view of the topic.

Assuming the relations of evaluation methods applied found by Klimashevskaia et al. can roughly be applied to other biases, our exclusion criteria demanding live or simulation based evaluation of their approach excludes around 90 % of studies [25]. This indicates either a large gap between evaluations of popularity bias and the broader RS and feedback loop scope, or limits our work.

To also consider future studies, our taxonomy might need to be extended to accurately capture additional mitigation types. Specifically, the general use of the learning problem category for reinforcement learning can be further refined.

The lack of widely applied simulation frameworks, such as a flexible benchmarking environment, indicates a potential direction for future research. This could increase comparability and further foster the already strong connection between academia and industry. Therefore, further assisting research should be directed at designing *responsible feedback loop systems*—ensuring both fairness to users, as well as the long-term health of recommender systems. Such efforts would guide academic exploration and facilitate practical adoption in real-world platforms.

REFERENCES

[1] Nil-Jana Akpinar, Cyrus DiCiccio, Preetam Nandy, and Kinjal Basu. 2022. Long-term Dynamics of Fairness Intervention in Connection Recommender Systems. en. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. doi: 10.1145/3514094.3534173.

[2] Len Bass, Qinghua Lu, Ingo Weber, and Liming Zhu. 2025. *Engineering AI Systems: Architecture and DevOps Essentials*. Addison-Wesley Professional.

[3] Christine Bauer, Eva Zangerle, and Alan Said. 2024. Exploring the landscape of recommender systems evaluation: practices and perspectives. en. *ACM Transactions on Recommender Systems*, 2, 1. doi: 10.1145/3629170.

[4] Christopher M Bishop. 2006. *Pattern recognition and machine learning*. Number 4. Vol. 4. Springer.

[5] Jennifer Brennan, Yahu Cong, Yiwei Yu, Lina Lin, Yajun Peng, Changping Meng, Ningren Han, Jean Pouget-Abadie, and David M. Holtz. 2025. Reducing Symbiosis Bias through Better A/B Tests of Recommendation Algorithms. en. In *Proceedings of the ACM on Web Conference 2025*. ACM. doi: 10.1145/3696410.3714738.

[6] Gökhan Çapan, İlker Gündoğdu, Ali Caner Türkmen, and Ali Taylan Cemgil. 2022. Dirichlet–Luce choice model for learning from interactions. en. *User Modeling and User-Adapted Interaction*, 32, 4. doi: 10.1007/s11257-022-09331-0.

[7] Simon Caton and Christian Haas. 2024. Fairness in machine learning: a survey. en. *ACM Comput. Surv.*, 56, 7. doi: 10.1145/3616865.

[8] Bo Chang et al. 2024. Cluster Anchor Regularization to Alleviate Popularity Bias in Recommender Systems. en. In *Companion Proceedings of the ACM Web Conference 2024*. ACM. doi: 10.1145/3589335.3648312.

[9] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: a survey and future directions. en. *ACM Trans. Inf. Syst.*, 41, 3. doi: 10.1145/3564284.

[10] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H. Chi. 2019. Top-K Off-Policy Correction for a REINFORCE Recommender System. en. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM. doi: 10.1145/3289600.3290999.

[11] Khalil Damak, Sami Khenissi, and Olfa Nasraoui. 2022. Debiasing the Cloze Task in Sequential Recommendation with Bidirectional Transformers. en. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM. doi: 10.1145/3534678.3539430.

[12] Andres Ferraro, Michael D. Ekstrand, and Christine Bauer. 2024. It's Not You, It's Me: The Impact of Choice Models and Ranking Strategies on Gender Imbalance in Music Recommendation. en. In *18th ACM Conference on Recommender Systems*. ACM. doi: 10.1145/3640457.3688163.

[13] Andres Ferraro, Dietmar Jannach, and Xavier Serra. 2020. Exploring Longitudinal Effects of Session-based Recommendations. en. In *Fourteenth ACM Conference on Recommender Systems*. ACM. doi: 10.1145/3383313.3412213.

[14] Dalin Guo, Sofia Ira Ktena, Pranay Kumar Myana, Ferenc Huszar, Wenzhe Shi, Alykhan Tejani, Michael Kneier, and Sourav Das. 2020. Deep Bayesian Bandits: Exploring in Online Personalized Recommendations. en. In *Fourteenth ACM Conference on Recommender Systems*. ACM. doi: 10.1145/3383313.3412214.

[15] Huifeng Guo, Jinkai Yu, Qing Liu, Ruiming Tang, and Yuzhou Zhang. 2019. PAL: a position-bias aware learning framework for CTR prediction in live recommender systems. en. In *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM. doi: 10.1145/3298689.3347033.

[16] Huizhong Guo, Zhu Sun, Dongxia Wang, Tianjun Wei, Jinfeng Li, and Jie Zhang. 2025. Enhancing new-item fairness in dynamic recommender systems. en. arXiv preprint arXiv:2504.21362. (2025). doi: 10.48550/arXiv.2504.21362.

[17] Priyanka Gupta, Ankit Sharma, Pankaj Malhotra, Lovekesh Vig, and Gautam Shroff. 2021. CauSeR: Causal Session-based Recommendations for Handling Popularity Bias. en. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. ACM. doi: 10.1145/3459637.3482071.

[18] Cuize Han, Pablo Castells, Parth Gupta, Xu Xu, and Vamsi Salaka. 2022. Addressing Cold Start in Product Search via Empirical Bayes. en. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. ACM. doi: 10.1145/3511808.3557066.

[19] Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. 2024. Bias mitigation for machine learning classifiers: a comprehensive survey. en. *ACM J. Responsib. Comput.*, 1, 2. doi: 10.1145/3631326.

[20] Chip Huyen. 2025. *AI Engineering: Building Applications with Foundation Models*. O'Reilly Media.

[21] Olivier Jeunen and Bart Goethals. 2023. Pessimistic Decision-Making for Recommender Systems. en. *ACM Transactions on Recommender Systems*, 1. doi: 10.1145/3568029.

[22] Olivier Jeunen and Bart Goethals. 2021. Pessimistic reward models for off-policy learning in recommendation. en. In *Proceedings of the 15th ACM Conference on Recommender Systems*. Association for Computing Machinery. doi: 10.1145/3460231.3474247.

[23] Sami Khenissi and Olfa Nasraoui. 2020. Modeling and counteracting exposure bias in recommender systems. (2020). https://arxiv.org/abs/2001.04832 arXiv: 2001.04832 [cs.IR].

[24] Anastasiia Klimashevskaia, Mehdi Elahi, Dietmar Jannach, Lars Skjærven, Astrid Tessem, and Christoph Trattner. 2023. Evaluating The Effects of Calibrated Popularity Bias Mitigation: A Field Study. en. In *Proceedings of the 17th ACM Conference on Recommender Systems*. ACM. doi: 10.1145/3604915.3610637.

[25] Anastasiia Klimashevskaia, Dietmar Jannach, Mehdi Elahi, and Christoph Trattner. 2024. A survey on popularity bias in recommender systems. en. *User Modeling and User-Adapted Interaction*, 34, 5. doi: 10.1007/s11257-024-09406-0.

[26] Yongkang Li, Xingyu Zhu, Yuheng Wu, Wenxu Zhao, and Xiaona Xia. 2025. A survey on causal inference-driven data bias optimization in recommendation systems: principles, opportunities and challenges. en. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15, 2. doi: 10.1002/widm.70020.

[27] Xiao Lin, Xiaokai Chen, Linfeng Song, Jingwei Liu, Biao Li, and Peng Jiang. 2023. Tree based Progressive Regression Model for Watch-Time Prediction in Short-video Recommendation. en. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM. doi: 10.1145/3580305.3599919.

[28] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Ji Yang, Minmin Chen, Jiaxi Tang, Lichan Hong, and Ed H. Chi. 2020. Off-policy Learning in Two-stage Recommender Systems. en. In *Proceedings of The Web Conference 2020*. ACM. doi: 10.1145/3366423.3380130.

[29] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. en. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, Virtual Event, Ireland. doi: 10.1145/3340531.3412152.

[30] Robert C Nickerson, Upkar Varshney, and Jan Muntermann. 2013. A method for taxonomy development and its application in information systems. en. *European Journal of Information Systems*, 22, 3. doi: 10.1057/ejis.2012.26.

[31] Nicolò Pagan, Joachim Baumann, Ezzat Elokda, Giulia De Pasquale, Saverio Bolognani, and Anikó Hannák. 2023. A classification of feedback loops and their relation to biases in automated decision-making systems. en. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM. doi: 10.1145/3617694.3623227.

[32] Yi Ren, Hongyan Tang, and Siwen Zhu. 2022. Unbiased Learning to Rank with Biased Continuous Feedback. en. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. ACM. doi: 10.1145/3511808.3557483.

[33] Yining She, Sumon Biswas, Christian Kästner, and Eunsuk Kang. 2025. FairSense: Long-Term Fairness Analysis of ML-Enabled Systems. en. (2025). doi: 10.48550/arXiv.2501.01665.

[34] Yi Su, Haokai Lu, Yuening Li, Liang Liu, Shuchao Bi, Ed H. Chi, and Minmin Chen. 2024. Multi-Task Neural Linear Bandit for Exploration in Recommender Systems. en. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM. doi: 10.1145/3637528.3671649.

[35] Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. en. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. Association for Computing Machinery. doi: 10.1145/3465416.3483305.

[36] Wei Tang, Chien-Ju Ho, and Yang Liu. 2021. Bandit Learning with Delayed Impact of Actions. en. doi: 10.5555/3540261.3542314.

[37] Xiangmeng Wang, Qian Li, Dianer Yu, and Guandong Xu. 2022. Off-policy Learning over Heterogeneous Information for Recommendation. en. In *Proceedings of the ACM Web Conference 2022*. ACM. doi: 10.1145/3485447.3512072.

[38] Tao Wu et al. 2020. Zero-Shot Heterogeneous Transfer Learning from Recommender Systems to Cold-Start Search Retrieval. en. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. ACM. doi: 10.1145/3340531.3412752.

[39] Hyunsik Yoo et al. 2024. Ensuring user-side fairness in dynamic recommender systems. en. In *Proceedings of the ACM Web Conference 2024*. Association for Computing Machinery. doi: 10.1145/3589334.3645536.

[40] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal intervention for leveraging popularity bias in recommendation. en. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery. doi: 10.1145/3404835.3462875.

[41] Yuqi Zhou, Sunhao Dai, Liang Pang, Gang Wang, Zhenhua Dong, Jun Xu, and Ji-Rong Wen. 2025. Exploring the escalation of source bias in user, data, and recommender system feedback loop. en. arXiv preprint arXiv:2405.17998. (2025). doi: 10.48550/arXiv.2405.17998.