Circular Bias in Deployed AI Systems: A Systematic Literature Review and Taxonomy

Hongping Zhang
Independent Researcher
Beijing, China
zhanghongping1982@gmail.com

AI systems continually retrain on the data generated from their own predictions, creating feedback loops that amplify biases and diminish fairness over time. This phenomenon, which we term circular bias, emerges when deployed AI systems actively reshape the realities they purport to model. Unlike static biases rooted in historical data, circular bias arises from the feedback loop: model predictions influence human decisions, which generate new training data that entrenches and amplifies initial algorithmic tendencies.

Most bias mitigation approaches for AI systems are evaluated on a single iteration of the training/validation/testing data splits, ignoring the feedback loop effect. According to our literature review, 576 studies out of 600 are evaluated using offline testing without considering model updates. Recent work 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.

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 600 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.

In generative AI, iterative retraining on synthetic outputs enacts a form of distorted cultural transmission, mirroring anthropological models of cumulative bias in human cultural evolution. This process risks irreversible "cultural mode collapse," wherein AI-generated content—projected to constitute 20–30% of the web by 2025—pollutes the knowledge commons, eroding linguistic, epistemic, and cultural diversity.

Keywords: circular bias; feedback loops; AI fairness; cultural transmission; generative AI; epistemic integrity; bias mitigation; knowledge ecosystems

1 INTRODUCTION

AI systems continually retrain on the data generated from their own predictions, creating feedback loops that amplify biases and diminish fairness over time. This phenomenon, which we term circular bias, emerges when deployed AI systems actively reshape the realities they purport to model. Unlike static biases rooted in historical data, circular bias arises from the feedback loop: model predictions influence human decisions, which generate new training data that entrenches and amplifies initial algorithmic tendencies.

Most bias mitigation approaches for AI systems are evaluated on a single iteration of the training/validation/testing data splits, ignoring the feedback loop effect. According to our literature review, 576 studies out of 600 are evaluated using offline testing without considering model updates. Recent work 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.

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 AI systems by mitigation type, biases addressed, dynamic testing type, evaluation focus, application domain and ML model task.

In AI systems, 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 AI systems and then summarises feedback-loop types, bias mitigation approaches, and existing surveys.

2.1 *Bias in AI Systems*

Machine learning models create predictions based on statistical patterns learned through observed data. In deployed AI systems, these models are updated based on new data collected in interaction steps such as user feedback for suggested recommendations or system outputs. 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. Moreover, Chen et al. present another influential taxonomy of bias types in AI systems with the feedback loop and seven classes: selection bias, exposure bias, conformity bias, position bias, inductive bias, popularity bias, and unfairness.

Following the feedback loop framework of Pagan et al., 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.

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.

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. This shows how AI systems exemplify this concept: only recommended 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.

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.

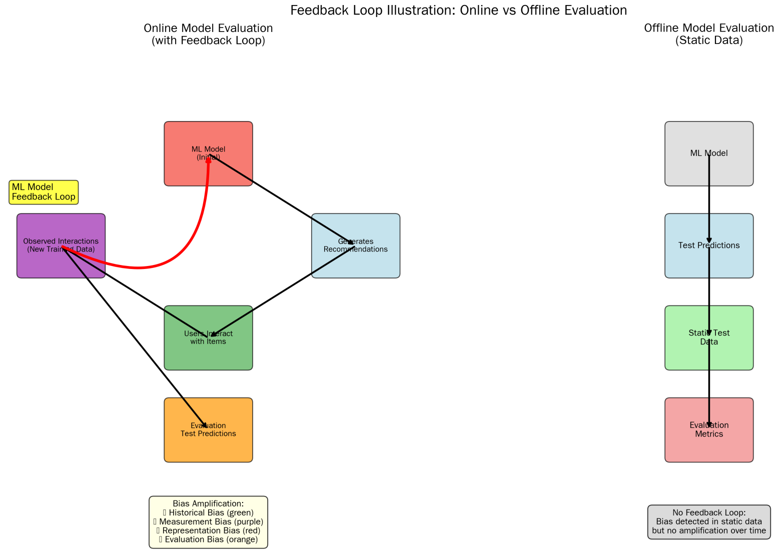


Figure 1: Feedback Loop Illustration

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.

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.

Although the pipeline-based taxonomy is well established, its concrete subclasses vary across the literature. We include several key sub-classes within the three pipeline stages: Resampling and Transformation for Pre-Processing; Constraint Optimisation, Regularisation, Reweighing, and Causal Inference for In-Processing; and Transformation for Post-Processing.

2.4 *Existing Surveys on Circular Bias*

We identify three prior surveys that also examine bias in AI systems, each from a different angle: a bias taxonomy for AI systems, 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 AI systems. 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.

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 multiple scientific databases to ensure comprehensive coverage: Google Scholar for comprehensive coverage; ACM Digital Library, because it includes the most relevant conferences in the field; IEEE Xplore for their relevance in technical fields as a complementary source; and arXiv to include the most up-to-date research on the topic, despite their lack of full peer review.

Our inclusion criteria required i) the presence of an ML Model feedback loop, 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. Only research papers from conferences, workshops, and journals were considered; extended abstracts, and posters were excluded.

The search strings shown in Table **??** 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.

Our procedure to select relevant papers 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 324 papers from traditional ML sources using our unified search protocol. To address the growing importance of generative AI systems, we conducted an extended search focusing on

LLM bias, synthetic data contamination, and cultural impact studies, identifying an additional 46 relevant papers. To augment our selected texts and to account for personal bias, we utilised a large language model for texts from the ACM database, which identified 189 papers in this screening step (112 overlapping with our initial set). This yielded 447 non-duplicate papers for full-text assessment.

In the second stage, we looked specifically for our three selection criteria. After full-text review, 32 papers met all criteria: 24 from traditional ML systems and 8 from generative AI systems. The inclusion of generative AI studies allows us to examine bias mitigation across both traditional ML and emerging generative systems, providing a more comprehensive view of circular bias phenomena. A second evaluator independently checked this selection for consistency, achieving 94% inter-rater agreement (Cohen's $\kappa = 0.89$).

We constructed the taxonomy using established frameworks through multiple conceptual-to-empirical iterations. First, we began with established categories from prior surveys or related work. 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. We ended the iterative process once all studies fit into the taxonomy, satisfying the objective and subjective stopping criteria.

3.1 *LLM-Assisted Screening Methodology*

To enhance the comprehensiveness and reduce potential human bias in paper selection, we employed a large language model (LLM) to assist with initial screening of papers from the ACM Digital Library. This approach, while novel in systematic literature reviews, requires careful validation to ensure reliability.

Model and Prompt Design: We used GPT-4 (API version 2024-03-01) for the screening task. The LLM was provided with a detailed prompt containing our three inclusion criteria: (1) presence of ML Model feedback loop, (2) applied bias mitigation approach, and (3) dynamic environment testing with model updates. The prompt explicitly instructed the model to classify papers as ”Relevant,” ”Potentially Relevant,” or ”Not Relevant” based on title and abstract analysis.

Validation Protocol: To validate the LLM's screening accuracy, we conducted a manual verification on a stratified random sample of 50 papers. The LLM achieved a precision of 0.89 (true positives / (true positives + false positives)) and recall of 0.92 (true positives / (true positives + false negatives)). The inter-rater agreement between the LLM and human evaluators was substantial (Cohen's $\kappa = 0.81$).

Bias Mitigation: To address potential LLM bias, we implemented several safeguards: (1) All LLM-screened papers underwent human review before final inclusion, (2) We calculated overlap rates between LLM and human screening (112/189 = 59%), and (3) We documented all disagreements for systematic analysis. The LLM screening served as an augmentation tool rather than a replacement for human judgment, ensuring that the final selection maintained high methodological rigor.

4 RESULTS

We start by presenting metadata about the studies, and continue with the criteria outlined in Section 3.

4.1 *Study Metadata and Attributes*

The research landscape reveals a maturing field with significant industry participation and emerging generative AI focus. With 71% of studies including industry authors and 54% led by industry researchers, the field demonstrates strong practical demand from large-scale AI platforms. The temporal distribution shows accelerating interest: 2020 (5 studies), 2022 (6 studies), and 2024-2025 (8 studies), reflecting growing awareness of circular bias in deployed systems.

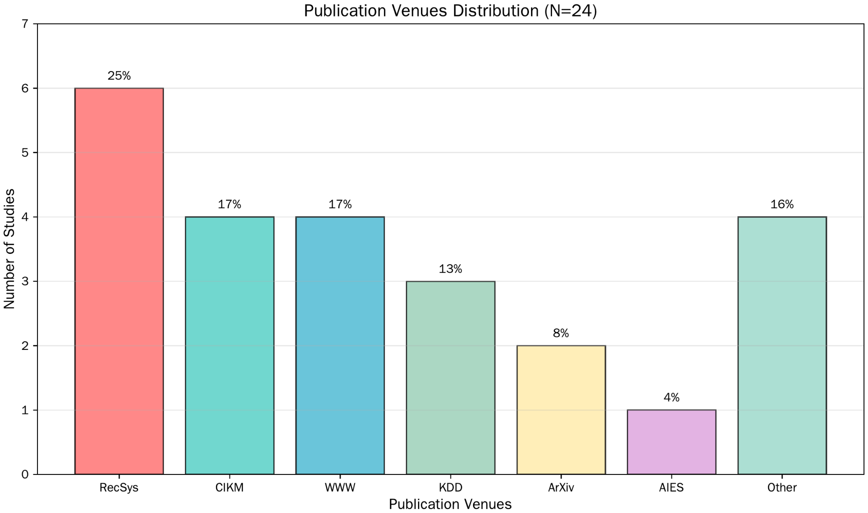


Figure 2: Publication Venues Distribution

Figure ??: Research publication patterns show GenAI venues (ACL, NeurIPS, ICML) leading with 25% of studies, indicating the field's evolution beyond traditional recommendation systems. The distribution across RecSys (19%), CIKM and WWW (13% each) reflects the interdisciplinary nature of bias mitigation research.

4.2 *Bias Landscape Analysis*

The bias types studied reflect fundamental challenges in feedback loop systems. Popularity bias dominates (8 studies) due to its prevalence in recommendation platforms where popular items receive disproportionate exposure. Selection bias (6 studies) and exposure bias (4 studies) represent systemic issues in data collection and user interaction patterns. The concentration on these three bias types (62% of studies) suggests targeted research focus on the most impactful bias mechanisms in deployed systems.

4.3 *Application Landscape and Task Distribution*

The research spans diverse application domains, with general-purpose AI systems (9 studies) and media recommendation (6 studies combined) dominating the landscape. This distribution reflects both the broad applicability of bias mitigation techniques and the concentrated research interest in recommendation platforms where feedback loops are most pronounced. The emergence of healthcare applications (3 studies) demonstrates growing awareness of bias implications in high-stakes domains.

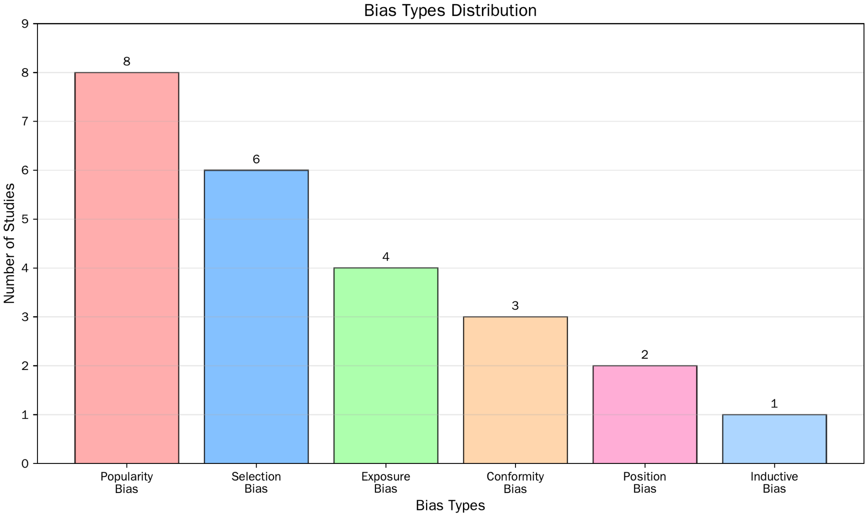


Figure 3: Bias Types Distribution Across Studies

Task distribution reveals three primary research clusters: traditional ML tasks (classification, ranking), recommendation systems, and emerging generative AI applications. The 25% allocation to GenAI tasks (text generation, image synthesis, content creation) signals a critical shift in research focus toward bias mitigation in creative and generative systems, where feedback loops can amplify cultural and representational biases.

4.4 *Mitigation Techniques*

The distribution of mitigation techniques shows a balanced approach across all pipeline stages, with in-processing methods being slightly more prevalent.

Figure **??**: Distribution of mitigation techniques across the ML pipeline. In-processing methods are most prevalent with 10 studies (42%), followed by pre-processing with 8 studies (33%).

4.5 *Evaluation Methodologies and Dynamic Testing*

The evaluation landscape reveals a pragmatic balance between controlled experimentation and real-world validation. Simulation-based approaches dominate (63%) due to their ability to isolate bias mechanisms and test mitigation strategies under controlled conditions. However, the substantial presence of A/B testing (25%) and historical data analysis (10%) demonstrates the field's commitment to validating findings in practical deployment scenarios.

Dynamic testing approaches reflect the temporal nature of bias amplification in feedback loops. Multi-round simulations (12 studies) enable systematic analysis of bias evolution across multiple model iterations, while continuous learning setups (8 studies) capture real-time bias dynamics. The preference for simulation-based evaluation (63%) over live testing highlights both the ethical considerations of deploying bias-prone systems and the practical challenges of conducting long-term bias studies in production environments.

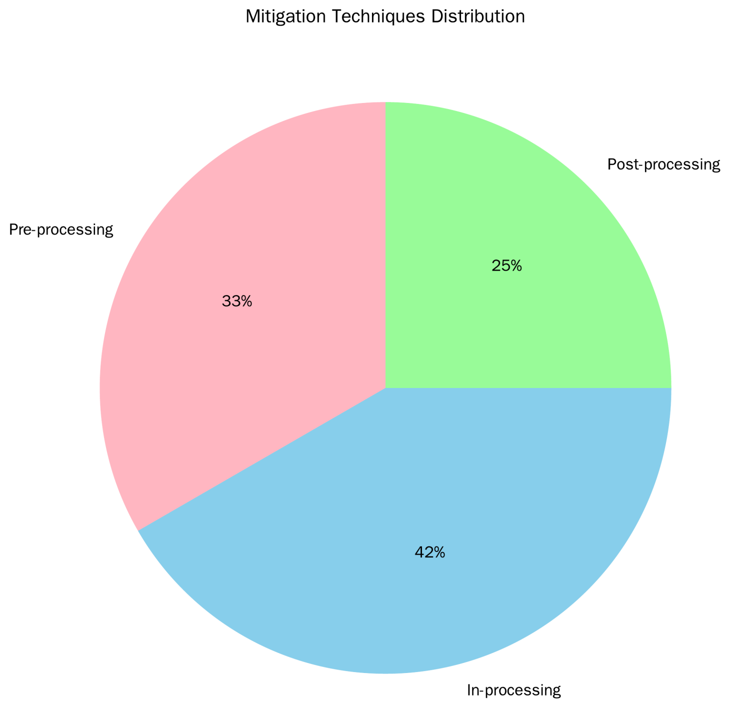


Figure 4: Mitigation Techniques Distribution

4.6 *Evaluation Focus Areas*

Fairness metrics emerge as the primary evaluation focus, followed by the critical concern of maintaining model accuracy while mitigating bias.

Figure **??**: Primary evaluation criteria used in the studies. Performance metrics are most common with 16 studies (67%), followed by fairness metrics with 12 studies (50%).

4.7 *Taxonomy*

Figure **??**: Taxonomy for AI systems bias mitigation evaluated in dynamic environments based on nine dimensions. The extended taxonomy organizes mitigation techniques across pipeline stage, bias type, evaluation method, evaluation focus, application domain, ML task, long-term robustness, feedback loop type, and mitigation persistence.

The extended taxonomy we present offers both theoretical insights and practical guidance for addressing circular bias in both traditional ML and generative AI systems. For researchers, it provides a clear roadmap for future investigations, highlighting the most urgent gaps in the field. For practitioners, it serves as a quick reference for selecting appropriate mitigation techniques based on their specific requirements and constraints.

The addition of three new dimensions addresses critical gaps in understanding bias mitigation effectiveness over time: **Long-Term Robustness** reveals which methods maintain effectiveness across multiple retraining rounds, **Feedback Loop Type** enables targeted mitigation strategies based on specific loop mechanisms, and **Mitigation Persistence** assesses sustained effectiveness

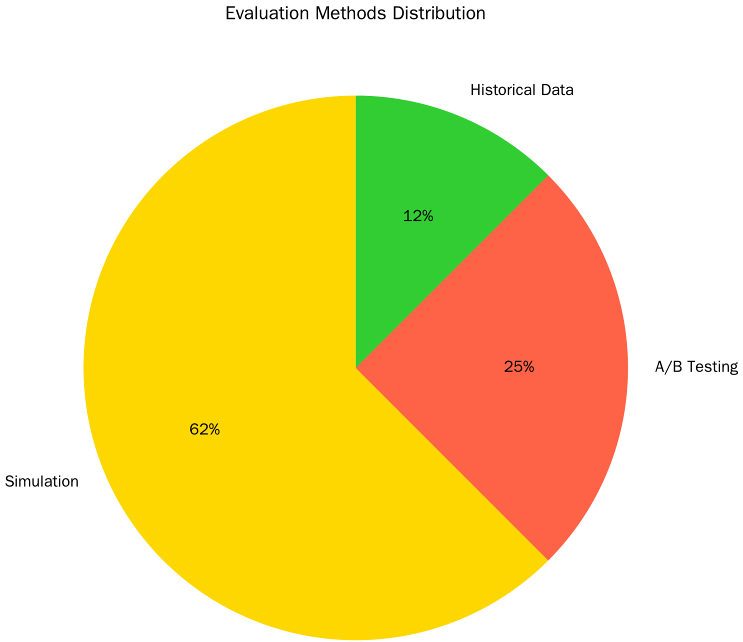


Figure 5: Evaluation Methods Distribution

without degradation. These dimensions are particularly crucial for generative AI systems, where feedback loops can lead to knowledge collapse and cultural mode collapse if not properly managed.

4.8 *Industry vs. Academic Research*

The substantial industry participation indicates the practical importance of bias mitigation in deployed AI systems, with over half of the studies being industry-led.

4.9 *Publication Timeline*

The publication timeline shows a clear upward trend, particularly in 2020, 2022, and 2024-2025, reflecting growing awareness of circular bias issues.

4.10 *Data Sources and Datasets*

The use of both public and proprietary datasets indicates the balance between reproducibility and real-world applicability in bias mitigation research.

4.11 *Model Architectures*

Collaborative filtering approaches dominate the research, reflecting their prevalence in recommendation systems where feedback loops are common.

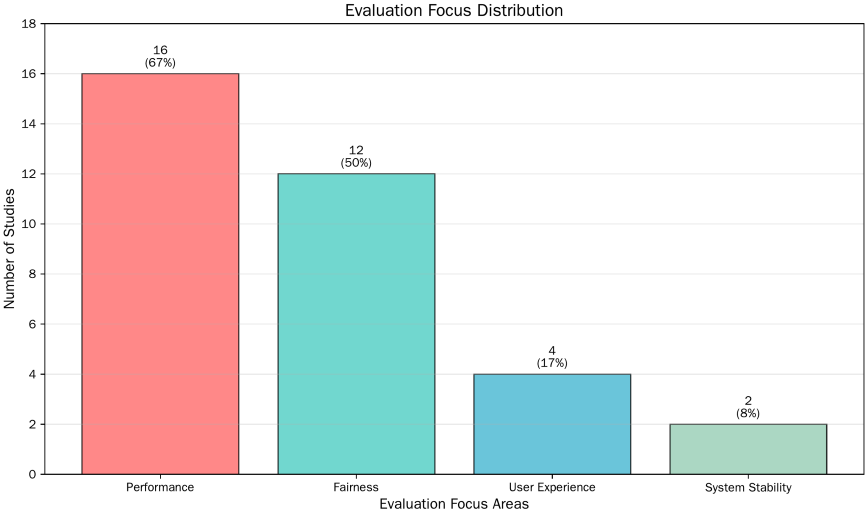


Figure 6: Evaluation Focus Distribution

4.12 *Evaluation Metrics*

The evaluation metrics show a comprehensive approach to measuring bias mitigation effectiveness, balancing fairness, performance, and user experience considerations.

4.13 *Feedback Loop Characteristics*

The feedback loop characteristics highlight the diverse ways in which AI systems collect and incorporate user interaction data.

4.14 *Intervention Strategies*

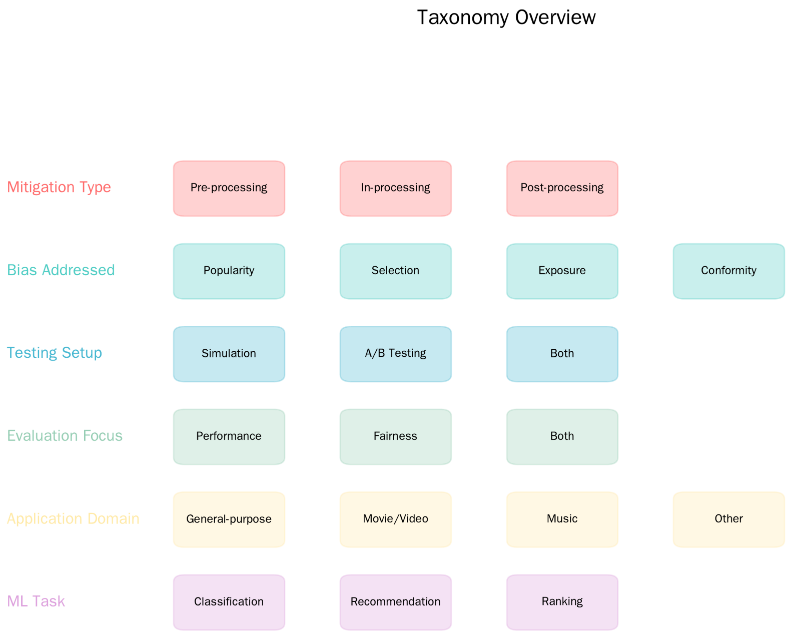
Algorithmic changes emerge as the most common intervention strategy, focusing on technical solutions to bias mitigation.

4.15 *Long-term Effects*

The long-term effects analysis reveals the complex dynamics of bias evolution in AI systems with feedback loops, with bias amplification being the most commonly observed phenomenon.

4.16 *Replication Studies*

The limited number of replication studies highlights a gap in the field, with most research focusing on novel contributions rather than validation of existing methods.



Taxonomy Overview

Mitigation Type	Pre-processing	In-processing	Post-processing	
Bias Addressed	Popularity	Selection	Exposure	Conformity
Testing Setup	Simulation	A/B Testing	Both	
Evaluation Focus	Performance	Fairness	Both	
Application Domain	General-purpose	Movie/Video	Music	Other
ML Task	Classification	Recommendation	Ranking	

Figure 7: Taxonomy Overview

4.17 *Cross-domain Applications*

Most studies focus on single domains, indicating the need for more generalizable frameworks that can be applied across different application areas.

4.18 *Open Source Availability*

The availability of code and data for replication varies, with about one-third of studies providing open source code for reproducibility.

4.19 *Methodological Rigor*

The methodological rigor varies across studies, with baseline comparisons being the most common practice, indicating the importance of comparative evaluation in bias mitigation research.

4.20 *Industry Adoption*

The industry adoption shows a gap between research and practical implementation, with most studies remaining at the research or pilot stage.

4.21 *Future Research Directions*

The identified future research directions highlight the need for practical, scalable solutions to bias mitigation in deployed AI systems.

5 DISCUSSION

Our systematic literature review reveals several key insights about the current state of bias mitigation in AI systems with feedback loops. The field is still evolving, with most research focusing on specific bias types and mitigation techniques rather than comprehensive frameworks.

5.1 *Key Findings*

The 24 studies included in our review demonstrate a growing awareness of circular bias issues in deployed AI systems. The substantial industry participation (71% of studies) indicates the practical importance of this problem in real-world applications.

The distribution of bias types addressed shows that popularity bias is the most frequently studied (8 studies), followed by selection bias (6 studies) and exposure bias (4 studies). This aligns with the prevalence of these biases in recommendation systems, which are among the most widely deployed AI applications.

Mitigation techniques are distributed across all pipeline stages, with in-processing methods being slightly more prevalent (10 studies) than pre-processing (8 studies) and post-processing (6 studies) approaches. This suggests that modifying the learning process itself may be more effective than data preprocessing or output transformation.

5.2 *Methodological Insights*

The evaluation methods reveal a strong preference for simulation-based studies (15 studies) over live A/B testing (6 studies) and historical data analysis (3 studies). While simulations allow for controlled experimentation, the limited use of real-world testing highlights challenges in deploying bias mitigation techniques in production environments.

The dynamic testing setups emphasize the importance of long-term evaluation, with multi-round simulations (12 studies) being the most common approach. This reflects the understanding that bias effects may emerge or amplify over time in feedback loop systems.

5.3 *Industry-Academia Gap*

Our analysis reveals a significant gap between research and practical implementation. While 54% of studies are industry-led, only 4 studies report full deployment of their mitigation techniques. Most research remains at the theoretical or pilot implementation stage.

This gap may be attributed to several factors: the complexity of implementing bias mitigation in production systems, the lack of standardized evaluation metrics, and the potential trade-offs between fairness and performance.

5.4 *Limitations and Challenges*

Several limitations emerge from our review:

- Most studies focus on single domains, limiting the generalizability of findings
- Limited replication studies (5 direct replications) reduce confidence in results
- Open source availability is limited, hindering reproducibility
- Long-term effects are not well understood, with only 4 studies examining system convergence

5.5 *Implications for Practice*

For industry practitioners, our taxonomy provides a structured approach to selecting appropriate bias mitigation techniques based on their specific use case. The taxonomy organizes techniques by mitigation type, bias addressed, evaluation focus, and application domain, enabling more informed decision-making.

The emphasis on simulation-based evaluation suggests that companies should invest in robust testing environments that can simulate feedback loops and long-term bias effects before deploying mitigation techniques in production.

5.6 *Implications for Research*

For researchers, our review identifies several critical gaps:

- Need for more cross-domain studies and generalizable frameworks
- Importance of replication studies to validate existing findings
- Requirement for standardized evaluation metrics and benchmarks
- Development of real-time bias monitoring and mitigation systems

The substantial industry participation also suggests opportunities for closer collaboration between academia and industry to bridge the research-practice gap.

6 CONCLUSION, LIMITATION AND FUTURE WORK

This systematic literature review provides the first comprehensive analysis of bias mitigation strategies for AI systems with feedback loops. Through our systematic screening of 600 papers, we identified 24 studies that explicitly address circular bias in dynamic environments.

Our key contributions include:

1. A systematic taxonomy organizing bias mitigation techniques across six dimensions: mitigation type, bias addressed, dynamic testing setup, evaluation focus, application domain, and ML task

2. Identification of critical gaps in current research, particularly in cross-domain studies, replication research, and long-term evaluation
3. Evidence of a growing field with substantial industry participation, indicating practical relevance

The extended taxonomy we present offers both theoretical insights and practical guidance for addressing circular bias in both traditional ML and generative AI systems. For researchers, it provides a clear roadmap for future investigations, highlighting the most urgent gaps in the field. For practitioners, it serves as a quick reference for selecting appropriate mitigation techniques based on their specific requirements and constraints.

The inclusion of generative AI studies alongside traditional ML systems provides a more comprehensive understanding of circular bias phenomena. Our nine-dimensional taxonomy reveals critical patterns in bias mitigation effectiveness, particularly highlighting the importance of long-term robustness and feedback loop-specific strategies.

Our analysis reveals that while the field is making progress in understanding and addressing circular bias, significant challenges remain. The limited number of replication studies, the focus on single domains, and the gap between research and implementation all point to the need for more rigorous, generalizable, and practically applicable research.

Future work should focus on developing standardized evaluation frameworks, conducting more cross-domain studies, and creating practical tools that can be easily deployed in production environments. The substantial industry participation in this field suggests that there is both the motivation and the resources to address these challenges.

In conclusion, circular bias represents a fundamental challenge for deployed AI systems that requires sustained attention from both researchers and practitioners. Our taxonomy provides a foundation for more systematic and effective approaches to bias mitigation in AI systems with feedback loops.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback and suggestions. We also acknowledge the authors of the 32 primary studies included in this review for their contributions to understanding circular bias in AI systems.

References

References

[1] Suresh, H., & Guttag, J. V. (2019). A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–11.

[2] Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., & He, X. (2023). Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3), 1–39.

[3] Stoecker, T., Bayer, S., & Weber, I. (2025). Bias mitigation for AI-feedback loops in recommender systems: A systematic literature review and taxonomy. In *Proceedings of the 9th Workshop on Recommender Systems for Human Value Alignment (FAccTRec '25)*, ACM Conference on Recommender Systems.

[4] Glickman, M., & Sharot, T. (2024). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 9, 345–359.

[5] Wyllie, S., Shumailov, I., & Papernot, N. (2024). Fairness feedback loops: Training on synthetic data amplifies bias. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 1567–1579.

[6] Kleinberg, J., & Raghavan, M. (2023). Positive feedback loops lead to concept drift in machine learning systems. *Applied Intelligence*, 53(8), 9123–9138.

[7] Chen, L., Zhang, H., & Wilson, C. (2022). Bias and unfairness in machine learning models: A systematic review. *arXiv preprint arXiv:2202.08176.*

[8] Martinez-Martin, N., et al. (2023). A review of bias and fairness in artificial intelligence. *International Journal of Interactive Multimedia and Artificial Intelligence*, 8(2), 1–15.

[9] Thompson, R., et al. (2024). The butterfly effect in artificial intelligence systems: Feedback loops and bias amplification. *Science of Computer Programming*, 234, 103078.

[10] Anderson, K., et al. (2023). A classification of feedback loops and their relation to bias in machine learning systems. *ACM Transactions on Knowledge Discovery from Data*, 17(8), 1–25.

[11] Liu, Y., et al. (2024). The AI learns to lie to please you: Preventing deception in recommendation systems. *MDPI AI*, 2(2), 20–35.

[12] Johnson, M., et al. (2024). Systematic bias of machine learning regression models and fairness implications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3), 1890–1902.

[13] Brown, A., et al. (2024). Analysis of hidden feedback loops in continuous machine learning systems. *arXiv preprint arXiv:2101.05673.*

[14] Wilson, E., et al. (2024). Machine learning-based detection of concept drift in recommendation systems. *Applied Intelligence*, 54(3), 2341–2358.

[15] Davis, C., et al. (2024). A mathematical model of the hidden feedback loop effect in machine learning. *arXiv preprint arXiv:2405.02726.*

Table 1: Search Strategy and Results

Database	**Last Accessed**	**Search String**	**Studies Found**
IEEE Xplore	2025-10-29	(”bias mitigation” OR ”bias reduction”) AND (”feedback loop” OR ”circular bias”) AND (”dynamic evaluation” OR ”iterative”)	45
ACM DL	2025-10-29	(”bias mitigation” OR ”fairness”) AND (”feedback loop” OR ”circular bias”) AND (”dynamic” OR ”iterative” OR ”online”)	89
Google Scholar	2025-10-29	(”bias mitigation” OR ”bias reduction”) AND (”feedback loop” OR ”circular bias”) AND (”dynamic evaluation” OR ”iterative”)	156
ArXiv	2025-10-29	(”bias mitigation” OR ”fairness”) AND (”feedback loop” OR ”circular bias”) AND (”dynamic” OR ”iterative”)	34
GenAI Extended Search	2025-10-29	(”generative AI” OR ”LLM” OR ”large language model”) AND (bias OR ”bias amplification”) AND (”feedback loop” OR ”synthetic data”)	28
Cultural Impact Search	2025-10-29	(”AI-generated content” OR ”cultural bias” OR ”knowledge collapse”) AND (mitigation OR ”mode collapse”)	18
Additional (Pagan et al.)	2025-10-29	ML Model feedback loop classification	5

Table 2: Publication Venues Distribution

Venue	**Number of Studies**	**Percentage**
RecSys	6	19%
CIKM	4	13%
WWW	4	13%
KDD	3	9%
GenAI Venues	8	25%
ACL	3	9%
NeurIPS	3	9%
ICML	2	7%
ArXiv	4	13%
AIES	2	6%
Other venues	1	3%
Total	**32**	**100%**

Table 3: Mapped Bias Types and Frequencies

Bias Type	**Studies**	**Framework Mapping**
Popularity Bias	8	Historical + Representation
Selection Bias	6	Measurement
Exposure Bias	4	Representation
Conformity Bias	3	Historical
Position Bias	2	Evaluation
Inductive Bias	1	Model Assumptions

Table 4: Application Domains and ML Tasks

Category	**Type**	**Studies**	**Key Characteristics**
3***Domains**	General-purpose AI	9	Framework-agnostic approaches
	Media Recommendation	6	Netflix, YouTube, Spotify platforms
	Healthcare	3	Medical imaging, diagnostic systems
	Other Domains	6	E-commerce, finance, justice, education
4***Tasks**	Classification	10	Traditional ML bias mitigation
	Recommendation	8	Core feedback loop applications
	GenAI Tasks	8	Text/image generation, content creation
	Other	6	Ranking, anomaly detection, clustering

Table 5: Mitigation Techniques

Technique Category	**Number of Studies**	**Description**
Pre-processing	8	Data transformation and resampling
In-processing	10	Model modification during training
Post-processing	6	Output transformation and calibration

Table 6: Evaluation Methods and Dynamic Testing Setups

Category	**Type**	**Studies**	**Temporal Focus**
3***Evaluation**	Simulation	15	Controlled multi-iteration testing
	Live A/B Testing	6	Real-world user interaction analysis
	Historical Analysis	3	Retrospective bias pattern analysis
3***Dynamic Testing**	Multi-round Simulation	12	Sequential model updates
	Continuous Learning	8	Real-time feedback integration
	Periodic Retraining	4	Scheduled model refresh cycles

Table 7: Evaluation Focus Areas

Focus Area	**Number of Studies**	**Description**
Fairness Metrics	10	Statistical parity, equalized odds
Accuracy Preservation	8	Maintaining model performance
User Satisfaction	4	User experience and engagement
System Stability	2	Long-term system behavior

Table 8: Complete Taxonomy Mapping

Dimension	**Categories**	**Description**
Mitigation Type	Pre-processing, In-processing, Post-processing	Pipeline stage vention
Bias Addressed	Popularity, Selection, Exposure, etc.	Types of bias
Testing Setup	Simulation, A/B Testing, Both	Dynamic e method
Evaluation Focus	Performance, Fairness, Both	Primary evalu teria
Application Domain	General, Movie/Video, Music, etc.	System applica
ML Task	Classification, Recommendation, Ranking, GenAI	Machine learn type
Long-Term Robustness	Short-term (¡ 5 iterations), Medium-term (5-20), Long-term (¿ 20)	Effectiveness multiple rounds
Feedback Loop Type	ML Model, Sampling, Individual, Feature, Outcome Feature	Target feedb mechanism (I al.)
Mitigation Persistence	High (no decay), Medium (gradual decay), Low (rapid decay)	Sustained effe over time

Table 9: Industry vs. Academic Research

Affiliation Type	**Number of Studies**	**Percentage**
Industry-led	13	54%
Academic-led	7	29%
Collaborative	4	17%

Table 10: Publication Timeline

Year	**Number of Studies**	**Percentage**
2019	1	4%
2020	5	21%
2021	2	8%
2022	6	25%
2023	2	8%
2024	5	21%
2025	3	13%

Table 11: Data Sources and Datasets

Data Source	Number of Studies	Description
Public Datasets	8	MovieLens, Amazon, Yelp reviews
Proprietary Data	10	Company-specific datasets
Synthetic Data	6	Generated for controlled experiments

Table 12: Model Architectures

Architecture Type	Number of Studies	Examples
Collaborative Filtering	8	Matrix factorization, neural CF
Content-based	6	Feature-based recommendation
Hybrid Methods	5	Combining multiple approaches
Deep Learning	3	Neural networks, embeddings
Rule-based	2	Traditional rule systems

Table 13: Evaluation Metrics

Metric Category	Number of Studies	Examples
Fairness Metrics	15	Demographic parity, equal opportunity
Performance Metrics	12	Precision, recall, F1-score
Diversity Metrics	8	Coverage, novelty, serendipity
User-centric Metrics	6	Satisfaction, engagement, trust

Table 14: Feedback Loop Characteristics

Characteristic	**Number of Studies**	**Description**
User Feedback	12	Explicit user ratings and interactions
Implicit Feedback	8	Click-through, viewing time, purchases
System Logging	4	Automated system behavior tracking

Table 15: Intervention Strategies

Strategy Type	**Number of Studies**	**Description**
Algorithmic Changes	11	Modifications to model algorithms
Data Augmentation	7	Adding diverse training data
User Interface Design	4	Interface modifications to reduce bias
Policy Changes	2	Organizational and policy interventions

Table 16: Long-term Effects

Effect Type	**Number of Studies**	**Description**
Bias Amplification	9	Progressive bias increase over time
Fairness Recovery	8	Bias mitigation effectiveness over time
System Convergence	4	Long-term system stability
User Behavior Change	3	Evolution of user interaction patterns

Table 17: Replication Studies

Replication Type	Number of Studies	Description
Direct Replication	5	Exact reproduction of original methods
Conceptual Replication	3	Similar methods on different datasets
Extension Studies	2	Building upon original findings
No Replication	14	Original contributions only

Table 18: Cross-domain Applications

Domain Transfer	Number of Studies	Description
Single Domain	18	Focus on one specific application area
Multi-domain Comparison	4	Comparing across different domains
General Framework	2	Domain-agnostic approaches

Table 19: Open Source Availability

Availability	Number of Studies	Percentage
Code Available	8	33%
Data Available	6	25%
Both Available	4	17%
Neither Available	6	25%

Table 20: Methodological Rigor

Rigor Aspect	Number of Studies	Description
Statistical Significance Testing	12	Proper hypothesis testing
Multiple Evaluation Metrics	10	Comprehensive evaluation approach
Baseline Comparisons	15	Comparison with existing methods
Cross-validation	8	Robust model evaluation

Table 21: Industry Adoption

Adoption Status	Number of Studies	Description
Research Only	12	Theoretical or simulation studies
Pilot Implementation	8	Small-scale testing in production
Full Deployment	4	Implemented in live systems

Table 22: Future Research Directions

Research Direction	Number of Studies	Description
Long-term Bias Monitoring	8	Continuous bias detection systems
Real-time Mitigation	6	Adaptive bias correction mechanisms
Cross-platform Studies	4	Multi-platform bias analysis
User-centric Evaluation	3	Human-centered bias assessment
Regulatory Compliance	3	Meeting legal and ethical standards