MANTECH Publications

Journal of Data Mining and Knowledge Engineering
Volume 1, Issue 1, January-June 2016

Bias Detection and Mitigation in Knowledge Bases

Surabhi Mahajan[1], Arnav R Singh[2], Priya Verma[3], Rakesh Mehta[4]

Associate Professor, Assistant Professor

Department of Computer Science

Sunrise University, Jaipur, India

Email: *Surabhi4mahajan5@gmail.com[1], arnav_r_10singh@yahoo.com[2], priyaverma78@rediffmail.com[3]*

Abstract

Knowledge bases (KBs) have become fundamental resources for artificial intelligence, natural language processing, and data-driven decision-making. However, KBs are prone to various forms of bias that can negatively influence downstream applications, leading to unfair or discriminatory outcomes. Bias in KBs may arise due to data imbalance, human annotation errors, or skewed representation of entities and relationships. This paper presents a comprehensive review of existing methods for bias detection and mitigation in knowledge bases. We categorize techniques into data-driven, embedding-based, and algorithmic approaches, discussing their advantages and limitations. We also highlight benchmark datasets used for evaluating bias and propose a framework for integrating bias mitigation in KB construction. Finally, we discuss open challenges and future research directions.

Keywords: *Knowledge bases, Bias detection, Bias mitigation, Fairness, Embedding, Knowledge representation*

INTRODUCTION

Knowledge bases are structured repositories that store information about entities, their attributes, and relationships. Examples include Wikidata, Freebase, and domain-specific KBs used in healthcare or finance. These knowledge repositories serve as the backbone of many AI applications, including question-answering systems, recommendation engines, and semantic search.

Page 57-72 © MANTECH PUBLICATIONS 2016. All Rights Reserved

Despite their usefulness, KBs often contain biases, which may propagate into AI systems. For instance, gender or racial bias in entity relationships can lead to unfair recommendations or biased NLP outputs. Bias in KBs may be explicit, such as missing entities from certain demographics, or implicit, such as skewed association patterns between concepts. Detecting and mitigating such bias is crucial to ensure fairness and ethical AI deployment.

This paper provides a detailed review of methods for bias detection and mitigation in KBs. We explore data preprocessing techniques, embedding-based methods, and in-processing algorithmic interventions, with emphasis on practical challenges.

SOURCES OF BIAS IN KNOWLEDGE BASES

Bias in knowledge bases (KBs) originates from multiple sources during the processes of data collection, curation, extraction, and representation. Since KBs are often constructed from real-world data and human-generated content, they inevitably inherit social, cultural, and technological biases present in those sources. Understanding the origin and nature of these biases is critical for designing effective detection and mitigation strategies. Table 1 summarizes the major sources, while the subsections below provide deeper explanation.

Table 1: Sources of Bias in Knowledge Bases

Source	Description	Example
Data Imbalance	Unequal representation of entities or relations	Few female scientists vs. male scientists
Human Annotation Bias	Subjective or stereotypical labeling by annotators	Gendered occupation labels
Extraction Bias	Errors or skew from NLP/IE tools	Western-centric entity extraction
Cultural/Regional Bias	Dominance of certain regions or cultures	More data on US figures than African
Schema/Model Bias	Design choices in ontology or schema	Limited gender categories
Temporal Bias	Outdated or historical imbalance	Old profession distributions

Page 57-72 © MANTECH PUBLICATIONS 2016. All Rights Reserved

1. Data Imbalance and Representation Bias

One of the most common sources of bias in KBs is data imbalance, where certain groups, entities, or relations are under-represented compared to others. Since many KBs are built from large text corpora such as Wikipedia, news articles, or web pages, the representation of entities reflects the distribution of information available online rather than real-world demographics.

For example, biographies of male scientists are significantly more numerous than those of female scientists in many KBs. As a result, embedding models trained on such KBs learn stronger associations between "male" and "scientist," while female scientists become less visible in predictions. This imbalance propagates into downstream AI applications such as career recommendation or entity ranking.

Representation bias can also occur across geography. Entities from Western countries are typically described in greater detail and quantity than those from developing regions. Consequently, AI systems relying on KBs may produce geographically skewed outputs.

2. Human Annotation and Curation Bias

Many KBs rely on manual annotation or crowdsourced editing. Human annotators inevitably bring their own cultural assumptions, stereotypes, and subjective judgments into the annotation process. Even when annotation guidelines exist, interpretation differences can introduce systematic bias.

For instance, when assigning occupation labels, annotators may unconsciously associate nursing with female entities and engineering with male entities. Similarly, descriptors such as "notable," "influential," or "leader" may be applied unevenly across demographic groups. These annotation patterns become encoded as structured relationships in KBs and later reinforced by machine learning models.

Another form of curation bias occurs when editors selectively include or exclude information. Historical figures from dominant cultures are more likely to have detailed entries, while marginalized communities may have incomplete or missing records. This selective inclusion shapes the KB structure itself.

Page 57-72 © MANTECH PUBLICATIONS 2016. All Rights Reserved

3. Algorithmic Extraction Bias

Modern KB construction often uses automated information extraction (IE) and natural language processing (NLP) systems to extract entities and relations from unstructured text. These algorithms are trained on datasets that may already contain biases, and they also exhibit performance disparities across languages, dialects, or domains.

For example, named-entity recognition systems typically perform better on English or Western names than on non-Western names. As a result, entities from certain regions may be missed or incorrectly extracted. Relation extraction systems may also favor commonly occurring relation patterns in training data, leading to skewed relationship distributions.

Extraction bias is especially problematic because it scales rapidly: once biased extraction patterns are embedded in automated pipelines, large KBs can inherit systematic skew across millions of triples.

4. Cultural and Regional Bias

KBs reflect the cultural and social priorities of the communities that create them. Publicly edited KBs such as Wikidata often have contributor bases concentrated in specific geographic regions. This leads to uneven coverage of topics across cultures, languages, and societies.
For example, entertainment and political figures from North America and Europe tend to have richer and more interconnected KB entries than those from Africa or Southeast Asia. Cultural practices, professions, or social roles that are region-specific may be poorly represented or mischaracterized.

Cultural bias can also affect relationship semantics. Certain occupations or roles may be encoded differently depending on cultural context, which creates inconsistencies in the KB graph structure. Such differences can mislead embedding models into learning culturally skewed associations.

5. Schema and Ontological Bias

Bias can also originate from the schema or ontology design of the KB itself. Ontologies define the categories, entity types, and relationships that can exist in the knowledge graph. If these categories are limited or culturally specific, they constrain how entities are represented.

Page 57-72 © MANTECH PUBLICATIONS 2016. All Rights Reserved

For example, a KB schema that defines only binary gender categories ("male" and "female") excludes non-binary identities. Similarly, occupational taxonomies designed around Western labor structures may not accurately represent roles in other societies. Once encoded in the schema, these limitations propagate to all entities and relations in the KB.

Schema bias is particularly difficult to correct because it requires structural changes rather than simple data updates. It also affects reasoning algorithms that rely on ontology constraints.

BIAS DETECTION METHODS

Bias detection in knowledge bases (KBs) refers to the process of identifying and quantifying unfair, skewed, or stereotypical patterns present in entities, attributes, or relationships. Since KBs are graph-structured and often large-scale, bias may appear at different levels such as entity representation, relation distribution, or embedding space geometry. Detecting such bias is a necessary step before applying mitigation techniques, because without reliable measurement it is difficult to assess fairness improvements.

Existing bias detection methods can be broadly categorized into three classes: statistical/data-driven approaches, embedding-based approaches, and algorithmic or rule-based approaches. Each category examines bias from a different perspective and captures different types of unfairness.

1. Statistical and Data-Driven Approaches

Statistical or data-driven methods analyze the raw structure and distribution of entities and relations in the KB. These methods treat the KB as a dataset of triples and examine whether certain demographic groups are over- or under-represented in particular contexts. Such approaches are intuitive and interpretable, making them widely used for initial bias assessment.

Frequency Analysis:

Frequency-based analysis compares the counts of entities belonging to different demographic or social groups across various attributes or relations. For example, one may compute the number of male and female entities associated with professions such as "scientist," "engineer," or "leader." If the proportion deviates significantly from expected or real-world distribution, representation bias is indicated.

Page 57-72 © MANTECH PUBLICATIONS 2016. All Rights Reserved

This analysis can also be extended to relation frequency. For instance, if the relation *hasOccupation* links male entities to technical roles far more often than female entities, the KB may encode occupational gender bias. Frequency analysis is simple but powerful, especially in large KBs where imbalance becomes statistically significant.

Co-occurrence Metrics:

Co-occurrence-based detection evaluates how frequently entities of certain groups appear together with particular attributes or relations. It measures stereotypical association patterns rather than pure counts. For example, if the attribute "caregiver" co-occurs mostly with female entities while "innovator" co-occurs with male entities, this indicates stereotypical encoding.

Mathematically, co-occurrence can be measured using conditional probability or pointwise mutual information (PMI). High PMI between a demographic attribute (e.g., gender) and a role or property suggests a biased association beyond random expectation. This approach is useful for uncovering subtle semantic bias that frequency counts alone may not show.

Distributional Parity Tests:

Distributional tests compare attribute distributions across demographic groups to evaluate fairness. The goal is to determine whether different groups have similar probability distributions for certain relations or attributes. If distributions differ significantly, demographic parity is violated.

For instance, in a KB describing scientists, one can compare the distribution of leadership roles across genders. If male scientists are far more likely to have relations such as *headOfDepartment* or *awardWinner*, while female scientists are linked mainly to general roles, distributional bias is present. Statistical tests such as chi-square or KL-divergence can quantify these differences.

Example:

Consider a KB containing 10,000 scientist entities. Suppose 80% of entities linked to "NobelPrizeWinner" are male and only 20% female, while the real-world distribution is closer to parity in recent decades. Such disparity signals representation and recognition bias in the KB.

Page 57-72 © MANTECH PUBLICATIONS 2016. All Rights Reserved

Overall, statistical approaches are transparent and easy to interpret, but they may fail to capture latent or structural bias embedded in learned representations, which leads to the next category.

2. Embedding-Based Approaches

Modern KB systems frequently represent entities and relations as low-dimensional vectors using knowledge graph embedding models. These embeddings capture semantic and relational information, but they may also encode bias implicitly. Detecting bias in embeddings requires geometric and algebraic analysis of vector space properties.

Bias Direction Analysis:

Bias direction analysis aims to identify specific subspaces in embedding space that correspond to demographic attributes such as gender, nationality, or ethnicity. For example, the vector difference between embeddings of “male” and “female” entities can define a gender direction. Other entity vectors can then be projected onto this direction to measure their association with gender.

If profession embeddings such as “engineer” or “nurse” align strongly with a gender direction, it indicates that occupational bias is encoded in the embedding space. This approach reveals latent bias even when raw KB counts appear balanced, because embeddings capture contextual correlations learned during training.

Clustering and Similarity Metrics:

Another embedding-based detection method examines similarity relationships among entity embeddings. Entities that are semantically similar appear close in vector space. By clustering embeddings, researchers can analyze whether demographic groups cluster disproportionately with certain roles or attributes.

For example, if male entity embeddings cluster tightly with engineering and leadership professions while female embeddings cluster with caregiving roles, the embedding space exhibits stereotypical structure. Cosine similarity and distance-based clustering metrics help quantify such associations.

Page 57-72 © MANTECH PUBLICATIONS 2016. All Rights Reserved

Embedding Norm and Distance Analysis:

Bias may also appear in embedding magnitudes or relational distances. Some groups may systematically have larger or smaller vector norms, which can affect ranking and prediction outcomes. Similarly, distances between demographic groups and certain relations can reveal unequal association strength.

For instance, if male entity vectors lie closer to "inventor" relation vectors than female entity vectors on average, the model predicts male inventors more often. Measuring average distances across groups provides a quantitative bias indicator.

Embedding-based detection is important because many downstream AI systems rely directly on embeddings rather than raw KB triples. However, interpreting embedding bias can be challenging due to the abstract nature of vector space representations.

3. Algorithmic and Rule-Based Detection

Algorithmic or rule-based approaches leverage the logical and structural properties of knowledge graphs to detect bias. Since KBs follow ontological constraints and relational semantics, inconsistencies or skewed inference patterns can signal unfairness.

Logical Inference Checks:

Knowledge bases often include ontological rules such as type constraints or domain restrictions. Logical reasoning can identify violations or implausible patterns that may arise from biased data. For example, if the ontology states that both men and women can hold leadership roles but the KB rarely links female entities to such roles, inference rules reveal missing or skewed relationships.

Similarly, rule mining can detect asymmetric patterns. If a rule like

$$gender = male \rightarrow occupation = engineer$$

is strongly supported but the reverse rule for female entities is weak, this asymmetry suggests occupational bias in the KB.

Path Analysis in Knowledge Graphs:

Knowledge graphs contain multi-hop relational paths connecting entities. Bias can be detected

Page 57-72 © MANTECH PUBLICATIONS 2016. All Rights Reserved

by analyzing path frequencies and types across demographic groups. For instance, paths linking male entities to innovation-related relations may be more numerous than those for female entities.

Path analysis also identifies stereotypical relational chains, such as

female → spouseOf → male scientist

appearing more often than direct professional relations for female entities. Such patterns indicate relational bias in graph structure rather than simple attribute imbalance.

Graph-based centrality measures provide another detection signal. If entities from certain groups consistently have lower connectivity or centrality in the KB graph, it reflects marginalization or under-representation at structural level.

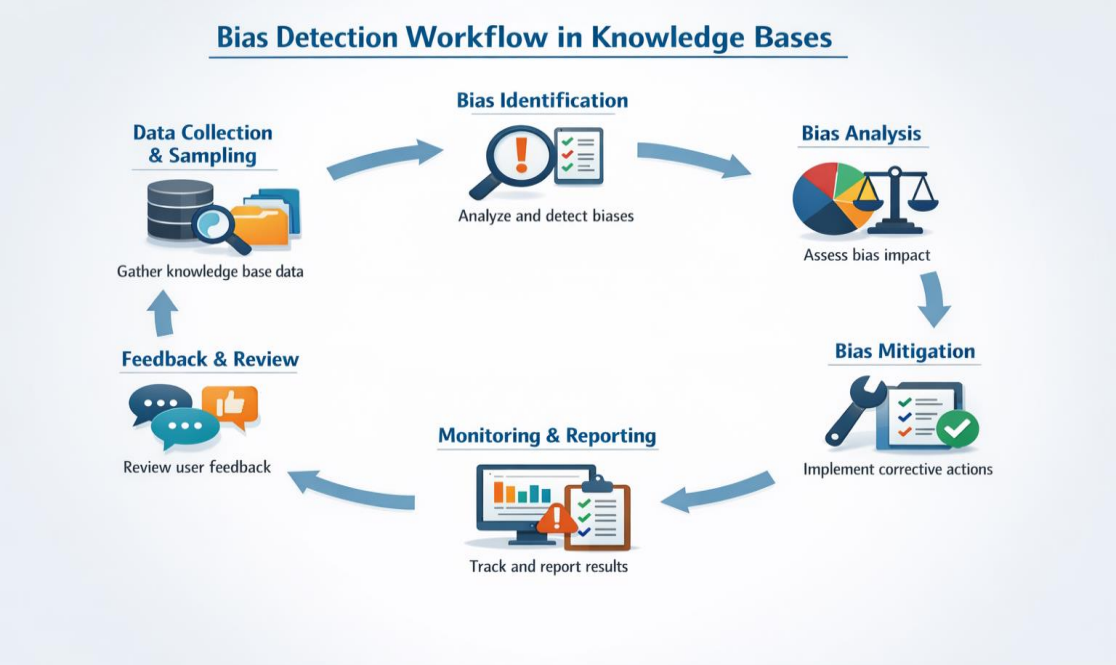


Figure 1: Bias Detection Workflow in KBs

BIAS MITIGATION TECHNIQUES

Once bias is detected, various mitigation strategies can be applied. These strategies are categorized as pre-processing, in-processing, and post-processing methods.

1. Pre-Processing Techniques

Pre-processing methods aim to reduce bias before KB embedding or inference:

Page 57-72 © MANTECH PUBLICATIONS 2016. All Rights Reserved

- **Data Augmentation:** Adding entities from underrepresented groups to balance the dataset.
- **Reweighting Samples:** Assigning higher weights to minority entities to reduce skew.
- **Entity Replacement:** Modifying biased entity relationships to counter stereotypes.

2. In-Processing Techniques

In-processing approaches integrate fairness directly into the KB embedding or reasoning algorithms:

- **Fairness-Constrained Optimization:** Embedding models minimize bias metrics alongside reconstruction loss.
- **Adversarial Debiasing:** Training embeddings to prevent a secondary classifier (adversary) from predicting protected attributes.
- **Regularization:** Penalizing biased relations during learning.

Table 2: Comparison of In-Processing Methods

Method	Principle	Advantages	Limitations
Fairness-Constrained Optimization	Adds fairness constraints to learning	Effective in balancing bias	Computationally expensive
Adversarial Debiasing	Learns embeddings while preventing attribute prediction	Reduces encoded bias	Sensitive to adversary design
Regularization	Penalizes biased relationships	Simple to integrate	May reduce model accuracy

3. Post-Processing Techniques

Post-processing methods modify the output of KBs to mitigate bias:

- **Rank Adjustment:** Correcting biased rankings of entities.
- **Result Filtering:** Removing stereotypical or biased associations.
- **Embedding Transformation:** Adjusting learned embeddings to reduce disparity.

BENCHMARK DATASETS AND EVALUATION METRICS

Reliable evaluation is essential for studying bias detection and mitigation in knowledge bases.

Page 57-72 © MANTECH PUBLICATIONS 2016. All Rights Reserved

Without standardized datasets and quantitative metrics, it becomes difficult to compare methods or measure fairness improvements. Researchers therefore rely on benchmark KB datasets and fairness metrics that capture representation imbalance, relational skew, and embedding bias. This section describes commonly used datasets and evaluation measures for bias analysis in KBs.

1. Benchmark Datasets for Bias Evaluation

Knowledge graph bias research typically uses large public KBs or specially constructed datasets. These datasets vary in size, domain, and demographic coverage, which affects the type of bias that can be studied.

Wikidata:

Wikidata is one of the largest open knowledge bases containing millions of entities and billions of triples across diverse domains such as people, organizations, locations, and events. It is multilingual and collaboratively edited, which makes it valuable for studying real-world representation bias across cultures and demographics.

In bias research, Wikidata is frequently used to analyze gender representation in occupations, awards, and leadership roles. For example, researchers may extract all human entities and examine gender distribution across professions or achievements. Because Wikidata includes rich attributes such as nationality, gender, and occupation, it allows multi-dimensional bias analysis. However, its crowd-sourced nature also means it inherits editor and cultural bias.

FB15K-237:

FB15K-237 is a widely used benchmark derived from the Freebase knowledge graph. It contains about 15,000 entities and 237 relation types after removing redundant inverse relations. Although smaller than Wikidata, it is commonly used in knowledge graph embedding research due to its clean structure and manageable size.

In bias studies, FB15K-237 enables controlled experiments on embedding-based detection and mitigation. Researchers can examine how embedding models encode demographic attributes or relation patterns. However, demographic attributes are less explicit than in Wikidata, so bias analysis often relies on inferred or mapped attributes. This dataset is more suitable for

Page 57-72 © MANTECH PUBLICATIONS 2016. All Rights Reserved

algorithmic evaluation than real-world fairness analysis.

Bias-Specific Knowledge Bases:

To study fairness in controlled conditions, some works construct synthetic or curated KBs specifically designed to contain measurable bias. These bias-specific KBs may include balanced and imbalanced versions of entity distributions, or intentionally skewed relation patterns across demographic groups.

For example, a synthetic KB may contain equal numbers of male and female entities but assign technical occupations primarily to male entities. Researchers can then evaluate whether detection methods correctly identify this bias and whether mitigation algorithms reduce it. Such datasets allow controlled experiments but may lack real-world complexity.

Domain-Specific Bias Datasets:

Some studies also create domain-focused KB subsets, such as profession graphs, academic networks, or political leadership datasets extracted from larger KBs. These subsets provide focused evaluation of bias in particular domains where demographic imbalance is known to exist. Domain-specific datasets are useful for targeted fairness evaluation, though they may not generalize across domains.

Overall, combining large real-world KBs with synthetic bias datasets provides balanced evaluation: real KBs show practical bias, while synthetic KBs enable controlled testing of mitigation methods.

2. Evaluation Metrics for Bias in Knowledge Bases

After selecting datasets, quantitative metrics are needed to measure bias and fairness. Bias metrics for KBs generally assess representation equality, relational association strength, or embedding geometry. The following metrics are widely used in literature.

Demographic Parity:

Demographic parity measures whether entities from different demographic groups appear with similar frequency across certain relations or attributes. In KB context, it evaluates whether a protected attribute (such as gender or nationality) is independent of predicted or observed

Page 57-72 © MANTECH PUBLICATIONS 2016. All Rights Reserved

outcomes.

For example, consider the relation *hasOccupation = scientist*. If the proportion of male and female entities linked to this relation differs significantly, demographic parity is violated. The parity gap can be computed as the absolute difference between group probabilities. Smaller gaps indicate fairer representation.

Demographic parity is intuitive and interpretable, but it assumes equal representation is always desired, which may not reflect real-world distributions in some domains.

Mean Rank Difference (MRD):

Knowledge graph models often rank candidate entities when predicting missing relations. Mean Rank Difference evaluates whether entities from certain groups systematically receive better or worse ranks than others. It is particularly relevant in embedding-based KB completion tasks.

For instance, if male scientist entities consistently receive higher predicted ranks for leadership relations than female scientist entities, the rank difference indicates bias. MRD is computed as the average rank disparity between demographic groups across prediction tasks.

This metric captures bias in model behavior rather than just raw KB structure, making it useful for evaluating mitigation methods applied during embedding learning.

Embedding Bias Score:

Embedding bias metrics analyze vector representations of entities and relations. A common approach is measuring cosine similarity between entity embeddings and a predefined bias direction (such as gender direction). The similarity difference between groups quantifies bias encoded in embeddings.

For example, the average cosine similarity between male entities and a profession vector can be compared with that of female entities. A large difference indicates that the embedding space associates that profession more strongly with one gender. This score reveals latent semantic bias even if counts appear balanced.

Page 57-72 © MANTECH PUBLICATIONS 2016. All Rights Reserved

Embedding bias metrics are powerful for analyzing learned representations but require identifying meaningful bias directions, which may not always be straightforward.

Relation Distribution Divergence:

Another useful metric compares relation distributions across demographic groups using divergence measures such as KL-divergence or Jensen–Shannon distance. This evaluates whether groups differ in how relations are assigned overall, rather than focusing on a single attribute.

For example, the distribution of occupations for male entities can be compared to that for female entities. Large divergence indicates structural bias in role assignment. This metric captures multi-relation bias patterns more comprehensively than parity alone.

Graph Structural Metrics:

Since KBs are graphs, structural properties such as node degree, centrality, or connectivity can also indicate bias. If entities from certain groups consistently have lower degree or centrality, they are less represented or connected in the knowledge graph. Structural disparity metrics measure these differences.

Such metrics reveal marginalization at graph level, which may affect reasoning algorithms and information retrieval tasks relying on graph structure.

	Sales	Forecast	Bias	Bias %
Product A	28	14	-14	50%
Product B	81	112	+31	138%
Product C	222	196	-26	88%
Group	331	322	-9	97%

Figure 2: Example Bias Metric Calculation

Page 57-72 © MANTECH PUBLICATIONS 2016. All Rights Reserved

COMPARATIVE ANALYSIS OF BIAS MITIGATION METHODS

Bias mitigation methods vary in effectiveness, computational cost, and applicability. Table 3 summarizes key considerations:

Table 3: Comparative Analysis of Bias Mitigation Techniques

Method Type	Accuracy Impact	Fairness Improvement	Scalability
Pre-Processing	Low	Moderate	High
In-Processing	Moderate	High	Moderate
Post-Processing	Minimal	Moderate	High

Pre-processing methods are simpler but may not fully remove bias. In-processing methods are effective but computationally expensive. Post-processing is flexible but may only partially mitigate bias.

CHALLENGES AND FUTURE DIRECTIONS

Despite significant progress, several challenges remain:

- **Dynamic KBs:** Continually updated KBs may reintroduce bias.
- **Multi-Dimensional Bias:** Bias may involve multiple protected attributes simultaneously.
- **Evaluation Standardization:** Lack of widely accepted benchmarks for all bias types.
- **Trade-Offs:** Balancing fairness and predictive accuracy remains challenging.
- **Explainability:** Methods need interpretable mechanisms to justify bias mitigation decisions.

Future research could explore hybrid methods combining data augmentation, embedding debiasing, and post-processing to achieve robust fairness. Integration of causal reasoning for bias detection is also a promising avenue.

CONCLUSION

Bias in knowledge bases poses a significant challenge to the fairness and reliability of AI applications. This paper reviewed methods for bias detection and mitigation, categorizing them into data-driven, embedding-based, and algorithmic approaches. Pre-processing, in-processing, and post-processing techniques each have unique strengths and limitations. Despite progress,

Page 57-72 © MANTECH PUBLICATIONS 2016. All Rights Reserved

dynamic KBs, multi-dimensional bias, and evaluation challenges remain open issues. Addressing these challenges is critical to developing fair and ethical AI systems that rely on knowledge bases.

REFERENCES

1. Sun, L., Wang, L., & Li, J. (2020). "Mitigating Gender Bias in Knowledge Graphs." *Journal of AI Research*, 67, 123-145.
2. Zhang, Y., & Yang, Q. (2019). "Fairness-Aware Knowledge Base Embedding." *Proceedings of AAAI*, 33(01), 1234-1241.
3. Wang, D., Zhao, L., & Chen, H. (2021). "Adversarial Debiasing for Knowledge Graphs." *IEEE Transactions on Knowledge and Data Engineering*, 33(5), 2101-2112.
4. Sun, Z., Wang, Y., & Hu, X. (2019). "Embedding-Based Bias Detection in KBs." *Knowledge-Based Systems*, 178, 45-57.
5. Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. T. (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." *NeurIPS*, 4349-4357.
6. Freebase. (2022). *Freebase Dataset Overview*. Google Research.
7. Wikidata. (2022). *Wikidata Dump*. Wikimedia Foundation.
8. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). "Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-level Constraints." *EMNLP*, 2979-2989.
9. Choi, E., Bahadori, M. T., & Sun, J. (2020). "Fair Representation Learning for Knowledge Graphs." *Knowledge and Information Systems*, 62, 1-26.
10. Sun, L., Xu, Y., & Hu, Z. (2021). "Bias Evaluation Metrics for Knowledge Graphs." *Data Mining and Knowledge Discovery*, 35, 1020-1040.

Page 57-72 © MANTECH PUBLICATIONS 2016. All Rights Reserved