Towards FAIR AI: A Survey of Trends and Knowledge Graph-Enhanced Bias Mitigation

Abhash Shrestha, Sanju Tiwari *Sr. Member, IEEE*, and Tek Raj Chhetri

***Abstract*—Artificial Intelligence (AI) is increasingly shaping daily life, from routine tasks to high-stakes applications such as healthcare diagnostics. Alongside these advances, concerns persist about the potential of AI systems to reinforce social inequities. While existing surveys have examined algorithmic bias in specific domains, there remains limited understanding of regional trends and the broader research landscape. This survey addresses that gap by reviewing 99 peer-reviewed articles to analyze patterns in AI-bias research. Our findings reveal significant disparities: developed countries—particularly the United States—dominate the field, whereas Asia, and especially the Global South, including India, are significantly underrepresented. Collaboration networks are also concentrated, with the United States and the United Kingdom leading, and most partnerships occurring within Europe, apart from notable United States collaborations with China and Hong Kong. We also identify a growing use of knowledge graphs as emerging tools for bias mitigation. To support ongoing research, we provide an interactive dashboard (https://cairnepal.github.io/algorithmic-bias-survey) that visualizes bias trends and regional collaborations in AI-bias research.**

***Index Terms*—Bias, Fairness, Knowledge Graphs, Regional, Fairness–Accuracy Trade-off**

I. Introduction

Today, the world has changed in ways we could never have imagined due to technological advancement, particularly artificial intelligence (AI). From influencing everything from everyday decisions, such as choosing personalized entertainment, to supporting critical operations in healthcare diagnostics [1], drug discovery [2], and manufacturing [3], AI continues to reshape our lives in unexpected yet profound ways. The winning of the 2024 Nobel Prize in Chemistry for the prediction and design of AI-driven protein structures signifies the ever-growing impact of AI on the world today [4].

However, despite great technological advancements in AI, limitations remain, the most notable of which is the issue of bias [5, 6, 7, 8, 9]. Fig. 1 illustrates the various types of bias in AI systems. It is critically important to address these AI biases. This is because biases in AI systems have direct and harmful consequences on human lives and society. For example, Zack et al. [8] conducted a study to understand the potential of GPT-4, a large language model (LLM), to perpetuate racial and gender bias in clinical diagnostics and recommendations. Zack et al. [8], in their study, found that GPT-4 does not model the demographic diversity of medical conditions, leading to biased responses. GPT-4 not only perpetuates existing biases, but also amplifies them, for example, by overrepresenting Asian and Hispanic populations in stereotypical conditions such as tuberculosis and being less likely to recommend advanced imaging procedures (e.g. CT, magnetic resonance imaging, or abdominal ultrasound) for black patients compared to white patients, thus raising concerns about its use for clinical decision support. Another prominent example is the case of COMPAS (Correctional Offender Management Profile for Alternative Sanctions), an AI-based tool employed in the US criminal justice system to assess the risk of recidivism (i.e., the likelihood of reoffending). Studies have revealed that COMPAS exhibits racial bias, often predicting a higher risk of reoffending for lack individuals compared to their white counterparts, even for similar profiles [10, 11]—directly affecting people's lives.

As these AI biases shape real-world social, economic, and health outcomes, they demand rigorous attention, and over the years many studies have been conducted that examined AI bias, its impacts, and mitigation techniques. For example, Ferrara [5] presents a systematic review of AI bias in healthcare, employment, criminal justice, and generative systems. In contrast, Ntoutsi et al. [6] examine bias primarily within the European context, framed their discussion around the EU (European Union) Charter of Fundamental Rights and the General Data Protection Regulation (GDPR). In addition, some investigations take a more targeted approach, for example, O'Connor and Liu's [12] study of gender bias, which explores the ways AI systems magnify gender disparities and surveys the corresponding mitigation strategies. Although existing studies [5, 6, 12, 13, 14] have made significant advances in understanding biases and mitigation strategies in various sectors, such as healthcare, more research is needed to explore these biases from different perspectives, including regional and sociocultural dimensions. This is essential, as what is considered normal or acceptable in one culture (or region) may not hold the same meaning or value in other cultures or regions around the world. For example, *in Western cultures, direct eye contact is often interpreted as a sign of confidence, politeness, honesty, attention, and positive engagement [15]. In contrast, in many Asian cultures, such as Japan, prolonged eye contact is often considered disrespectful [16].* AI (or AI systems) trained on datasets reflecting only one cultural (or region or country) norm may therefore misinterpret or misclassify behavior from other regions, leading to biased outcomes.

A key contributing factor is the limited participation in

Abhash Shrestha is with the Center for Artificial Intelligence (AI) Research Nepal, Sundarharaincha-09, Koshi, Nepal (e-mail: abhash.shrestha@cair-nepal.org).

Sanju Tiwari is with Sharda University, Delhi-NCR, India, and with Shodhguru Innovation and Research Labs, India (e-mail: tiwarisanju18@ieee.org).

Tek Raj Chhetri is with the Center for Artificial Intelligence (AI) Research Nepal, Sundarharaincha-09, Koshi, Nepal, and with the McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA (e-mail: tekraj.chhetri@cair-nepal.org; tekraj@mit.edu). Corresponding author.

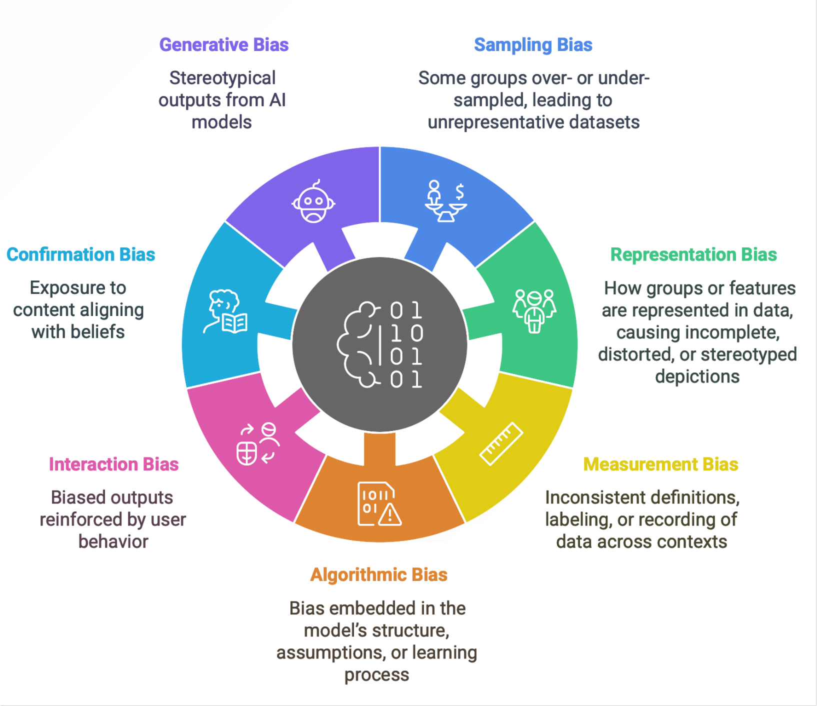

Generative Bias
Stereotypical outputs from AI models
Sampling Bias
Some groups over- or under-sampled, leading to unrepresentative datasets
Confirmation Bias
Exposure to content aligning with beliefs
Representation Bias
How groups or features are represented in data, causing incomplete, distorted, or stereotyped depictions
Interaction Bias
Biased outputs reinforced by user behavior
Measurement Bias
Inconsistent definitions, labeling, or recording of data across contexts
Algorithmic Bias
Bias embedded in the model's structure, assumptions, or learning process

Fig. 1. Biases in AI Systems

research from diverse regions, which can cause a lack of essential region-specific knowledge — including cultural and societal context — into the design and evaluation of AI systems. This lack of diversity skews both the data and the research landscape, resulting in systems and research that disproportionately reflect the perspectives of a few regions.

To address this gap, our study systematically surveys and analyzes the regional disparity in AI bias research, highlighting how certain regions are underrepresented in current scholarly efforts. Additionally, we provide an overview of how knowledge graphs (KGs) - widely regarded a key enabling technology to address bias in modern AI [17], including large language models (LLM) - can be used for bias mitigation. Moreover, we also categorize biases along the AI lifecycle, from data collection to user interaction.

KGs, a foundational technology of the Semantic Web, offer formalized representations where nodes denote entities of interest and edges define the relationships between them. Their ability to integrate domain-specific knowledge and provide rich contextual awareness makes them particularly valuable in addressing the limitations of current AI paradigms—especially large language models (LLMs)—which often lack such grounding and are prone to hallucinations [17, 18, 19, 20, 21]. Building on this perspective, our work surveys existing techniques that leverage KGs for bias mitigation, thereby addressing a key gap in the current research landscape.

In this survey, we systematically reviewed 99 peer-reviewed articles to address the following research questions (RQ).

- *RQ1: How does the distribution of AI bias research vary across different regions and what are the gaps in regional representation?*
- *RQ2: How has research in various domains evolved overtime?*
- *RQ3: How can Knowledge Graphs be leveraged as an effective tool for AI bias mitigation?*

The remainder of the paper is organized as follows. Section II reviews related work, while Section III outlines the research methodology. Section IV examines various types of biases in AI, and Section V discusses their impact on people. Section VI explores strategies for mitigating bias across different phases of the AI lifecycle. Section VII analyzes the current AI research and development landscape across countries and institutions, and Section VIII investigates graph-based approaches, including the use of ontologies and knowledge graphs, for bias mitigation. Section IX presents the discussion, Section X provides recommendations derived from our analysis, and Section XI concludes the paper with future directions.

II. Related Works

This section reviews related works, focusing on studies that survey bias and fairness. We have excluded papers that mention bias only in passing or narrowly address a single aspect (e.g., gender bias in a specific application or a single demographic group) without offering broader generalizations or methodological insights for mitigation.

Hort et al. [13] conducted a comprehensive survey of 341 publications on bias mitigation for machine learning (ML) classifiers, making it a valuable resource in this domain. Their work focuses on identifying bias mitigation practices, including the metrics used for evaluation and the applied methods. The study highlights key challenges, such as the inconsistency of evaluation metrics—with 109 different metrics reported across studies—and the limited availability of diverse datasets that reflect real-world scenarios. The survey also observed a sharp rise in bias mitigation research after 2018. Ferrara [5] conducted a systematic review studying AI bias and fairness. Their study specifically examines the societal impact of AI bias—including in generative AI—while also exploring mitigation strategies similar to those proposed by Hort et al. [13], and analyzing the limitations of current approaches such as group fairness. The work emphasizes key application domains including healthcare, employment, and criminal justice. Similarly, the study by Mehrabi et al. [14] presents a comprehensive survey of bias sources, fairness definitions, and mitigation techniques, offering a broad overview of challenges and approaches across multiple application domains.

Ntoutsi et al. [6] take a multidisciplinary approach to studying bias and fairness, incorporating perspectives from computer science, law, and social sciences, with particular emphasis on European Union (EU) regulatory frameworks. Additionally, this study also focuses on modeling bias in ontology, which is used in symbolic AI [22], that is omitted in other surveys, which primarily concentrate on statistical or machine learning-based methods.

Gurupur et al. [23] focus on bias in AI-based healthcare decision support systems, proposing a transdisciplinary approach to address two main bias types: knowledge bias (from flawed experiments or shallow expertise) and processing bias (from algorithm choice and feedback loops). They emphasize the need validate both outputs and evaluation tools. Similarly, Fan et al. [24] survey recommender systems from dimensions including fairness, explainability, and privacy. They categorize bias into data, model, and feedback loop bias, review mitigation strategies across the pipeline, and highlight knowledge graphs (KGs) as useful for both generating interpretable explanations and supporting privacy-preserving, bias-aware recommendations. Shahbazi et al. [25] similarly focus on representation bias, offering a structured taxonomy of its subtypes and outlining both algorithmic and data-level interventions.

Tang et al. [26] and Caton et al. [27] have performed comprehensive surveys on fairness and bias mitigation in AI, organizing methods into the standard pre-, in-, and post-processing framework. Caton et al. focus on supervised learning, especially binary classification, while Tang et al. follow a more theoretical, interdisciplinary approach, linking ML fairness to moral and political philosophy.

Research such as the ones done by Li et al., Gallegos et al., Chu et al., Want et al., Zhou et al., Dudy et al. [28, 29, 30, 31, 32, 33] focus on bias detection and mitigation from an LLM perspective. Chu et al. [30] have propose a granular taxonomy of fairness in LLMs where as Wang et al. [31] provide a detailed taxonomy of research on factuality in Large Language Models consisting of the issue, evaluation, analysis and enhancement. While Zhou et al. [32] focus on the development, application, evaluation, and bias mitigation of large language models (LLMs) within the medical domain, Dudy et al. [33] investigate geographical bias in LLMs across regions in the United States, analyzing how such models may discriminate against certain areas in terms of relocation advice, tourism recommendations, and business potential .

Although existing reviews have made substantial contributions to understanding bias in AI and have proposed various mitigation strategies, often incorporating legal and societal perspectives—they still fall short in addressing several critical dimensions. Specifically, (i) *understanding the regional differences in terms of funding, research domain, and contribution towards bias*; and (ii) *evolution of bias research in various domains across time i.e. a temporal analysis of the research landscape* g This study addresses this gap (see Table **??**) by providing a comprehensive overview of how bias manifests itself and varies in different geographical and cultural contexts.

III. Methodology

This section presents an overview of our methodology. Section III-A outlines the approach used for conducting the literature review, while Section III-B describes the methodology employed for categorizing the reviewed literature.

A. Literature Review Methodology

Fig. 2 shows the methodology used to compile the relevant literature for this survey. Our approach was designed to comprehensively identify relevant research related to bias in AI-driven decision-making systems. We began by formulating a set of search strings centered on key concepts such as bias, artificial intelligence, and decision-making, as illustrated in Fig. 2. These search strings were then used to conduct systematic queries across five major digital libraries: IEEE Xplore, ACM Digital Library, Scopus, ScienceDirect, and Engineering Village.

To ensure relevance and quality, we applied inclusion criteria including: (i) publication date from 2015 onward, (ii) peer-reviewed status, and (iii) explicit focus on bias within AI or machine learning contexts. We excluded papers that lacked sufficient methodological detail, were not written in English, had minimal citation impact, or were deemed irrelevant based on their titles and abstracts. Following this initial filtering, we conducted a manual review of each paper to confirm alignment with our research objectives. To further enhance coverage and reduce the risk of missing influential studies, we complemented the search with graph-based discovery tools such as Connected Papers[1] and Litmaps[2]. These tools enabled us to trace citation patterns and find relevant works potentially overlooked by keyword search alone. This integrated strategy - which combined keyword-based querying, manual screening, and citation network analysis - resulted in a curated corpus that underpins the discussions and insights presented in the remainder of this paper.

B. Systematic Categorization of Literature

The research papers reviewed in this study were broadly classified into five categories (see Fig. 3) according to their main focus: (1) *Health & Clinical AI*; (2) *Recommender Systems*; (3)

[1] https://www.connectedpapers.com

[2] https://www.litmaps.com

TABLE I
COMPARISON OF EXISTING SURVEYS AND OUR WORK

Survey Paper	**Bias Types**	**Fairness Methods**	**Regional Analysis**	**KG for Bias Mitigation**
Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey [13]	✓	✓	✗	✗
Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies [5]	✓	✓	x	x
Bias in data-driven artificial intelligence systems—An introductory survey [6]	✓	✓	x	x
Inherent Bias in Artificial Intelligence-Based Decision Support Systems for Healthcare [23]	x	✓	x	x
A Survey on Bias and Fairness in Machine Learning [14]	✓	x	✓	x
A Comprehensive Survey on Trustworthy Recommender Systems [24]	✓	✓	x	x
What-is and How-to for Fairness in Machine Learning: A Survey, Reflection, and Perspective [26]	x	✓	x	x
Representation Bias in Data: A Survey on Identification and Resolution Techniques [25]	x	✓	x	x
Fairness in Machine Learning: A Survey [27]	x	✓	x	x
A Survey on Fairness in Large Language Models [28]	✓	✓	x	x
Bias and Fairness in Large Language Models: A Survey [29]	✓	✓	x	x
Fairness in Large Language Models: A Taxonomic Survey [30]	✓	✓	x	x
Survey on Factuality in Large Language Models: Knowledge, Retrieval, and Domain-Specificity [31]	✓	x	x	x
A Survey of Large Language Models in Medicine: Progress, Application, and Challenge [32]	✓	✓	x	x
Unequal Opportunities: Examining the Bias in Geographical Recommendations by Large Language Models [33]	✓	x	✓	x
Towards improving prediction accuracy and user-level explainability using deep learning and knowledge graphs: A study on cassava disease [19]	✓	x	✓	x
A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts [34]	✓	x	✓	x
Our Work	✓	✓	✓	✓

LLMs & NLP; (4) *General Fairness & Bias Mitigation*; and (5) *Graph-Based Fairness & Bias Mitigation*. This categorization helped structure our analysis and highlight domain-specific challenges and mitigation strategies.

The first category, *Health & Clinical AI*, includes studies conducted in healthcare and clinical settings, focusing on tasks such as clinical decision support and other medical applications of AI, with an emphasis on identifying and mitigating biases that may affect patient outcomes. Similarly, the second, *Recommender Systems*, includes studies that predominantly focused on biases that can occur within recommender systems and how to mitigate them, while *LLMs & NLP*, the third category, examines the fairness concerns in LLMs and natural language processing systems, particularly those related to biased language understanding and generation. The fourth category, *General Fairness & Bias Mitigation*, comprises research focused on identifying and addressing bias across AI-powered systems, irrespective of the application domain. The fifth category, *Graph-Based Fairness & Bias Mitigation*, includes studies that investigate methods for detecting and mitigating bias in graph-based technologies, such as knowledge graphs and algorithms. However, it is important to note that some studies in our categorization may overlap across multiple categories, reflecting the interconnected nature of bias mitigation approaches in different types of systems. In such cases, those papers were included in all relevant categories.

IV. BIASES

Biases can arise at various stages of AI development, affecting fairness, reliability, and overall system performance. To effectively mitigate these biases, it is crucial to understand how they emerge and propagate throughout the AI lifecycle. Biases can arise at any point, as seen in Fig. 4, from data collection phase to the user interaction phase.

A. A brief overview of Bias and its types

a) Data Collection Bias: Data collection is a foundational phase in the development of AI or AI systems. Biases introduced at this stage can have a compounding effect, directly influencing downstream model performance and decision-making processes. Such biases typically arise from flawed sampling methodologies or demographic imbalances that fail to accurately capture real-world distributions [35, 36]. Three principal types of bias may emerge during data collection: sampling bias, representation bias, and measurement bias. Sampling bias occurs when certain groups are systematically over-or underrepresented in a dataset, resulting in skewed model performance [37]. Similarly, representation bias arises when underlying data distributions fail to capture the diversity of the target population, causing minority groups to be overlooked [25]. Measurement bias emerges when variables are recorded or labeled inconsistently, distorting results [38]. For instance, subjective ratings of "exercise intensity," in a

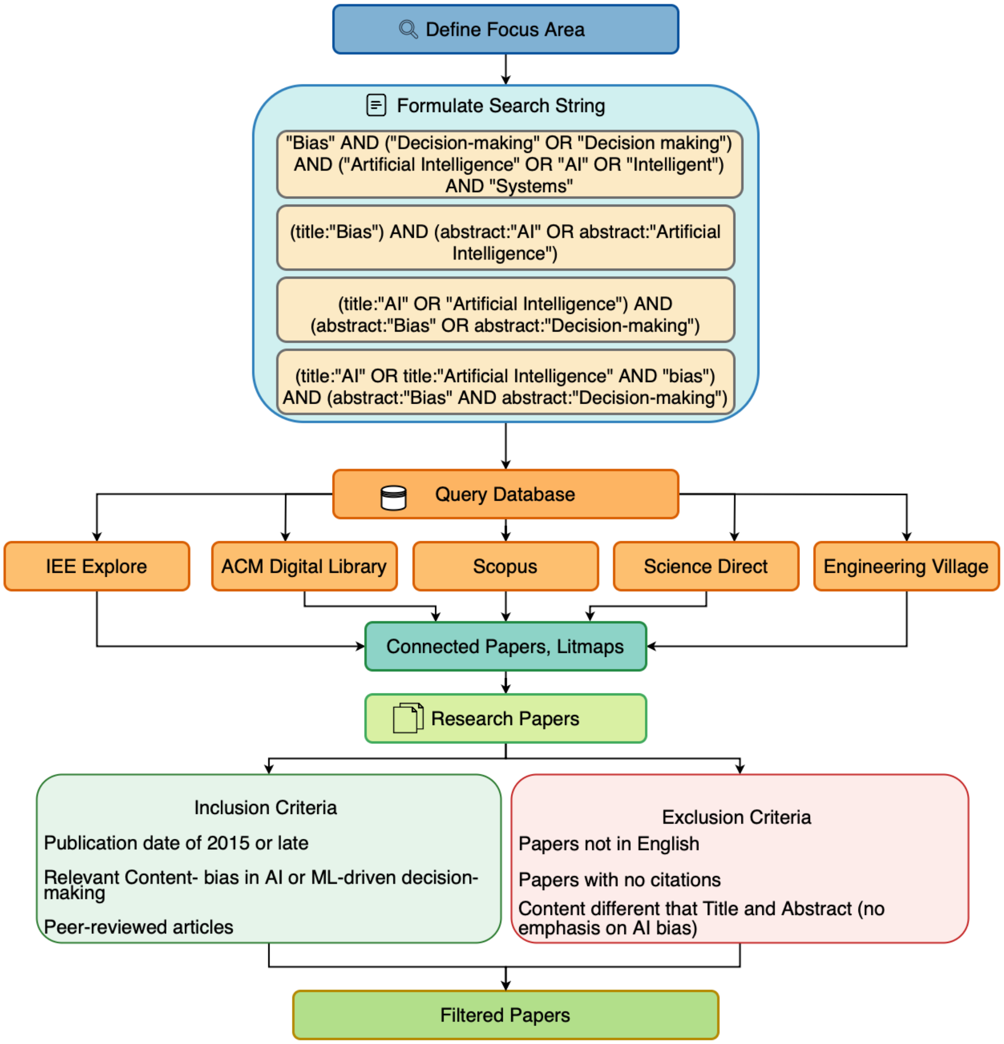

Define Focus Area
Formulate Search String
"Bias" AND ("Decision-making" OR "Decision making") AND ("Artificial Intelligence" OR "AI" OR "Intelligent") AND "Systems"
(title:"Bias") AND (abstract:"AI" OR abstract:"Artificial Intelligence")
(title:"AI" OR "Artificial Intelligence") AND (abstract:"Bias" OR abstract:"Decision-making")
(title:"AI" OR title:"Artificial Intelligence" AND "bias") AND (abstract:"Bias" AND abstract:"Decision-making")
Query Database
IEE Explore
ACM Digital Library
Scopus
Science Direct
Engineering Village
Connected Papers, Litmaps
Research Papers
Inclusion Criteria
Publication date of 2015 or late
Relevant Content- bias in AI or ML-driven decision-making
Peer-reviewed articles
Exclusion Criteria
Papers not in English
Papers with no citations
Content different that Title and Abstract (no emphasis on AI bias)
Filtered Papers

Fig. 2. Methodology overview for literature search and selection in this review.

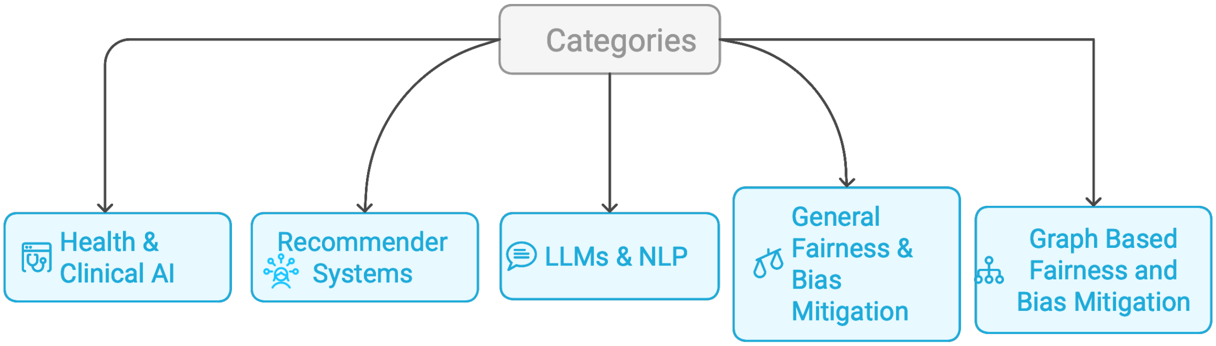

Categories
Health & Clinical AI
Recommender Systems
LLMs & NLP
General Fairness & Bias Mitigation
Graph Based Fairness and Bias Mitigation

Fig. 3. Overview of the literature classification into five thematic categories based on the focus of each study.

survey or respondents tailoring answers because they know the study's aims, are common examples [39, 40].

b) Bias in Algorithm Design: The data collection phase is followed by the model development phase in the AI lifecycle. This phase determines how the system learns from the data and makes predictions. During this phase, the most prevalent type of bias is algorithmic bias.

Algorithmic bias occurs when model architectures or optimization processes systematically disadvantage certain groups—even with balanced data-by reinforcing historical patterns, such as racial disparities in criminal justice (e.g., COMPAS).[35, 36, 41].

c) User Interaction Bias: This type of bias arises post-deployment when user behavior and overreliance on AI

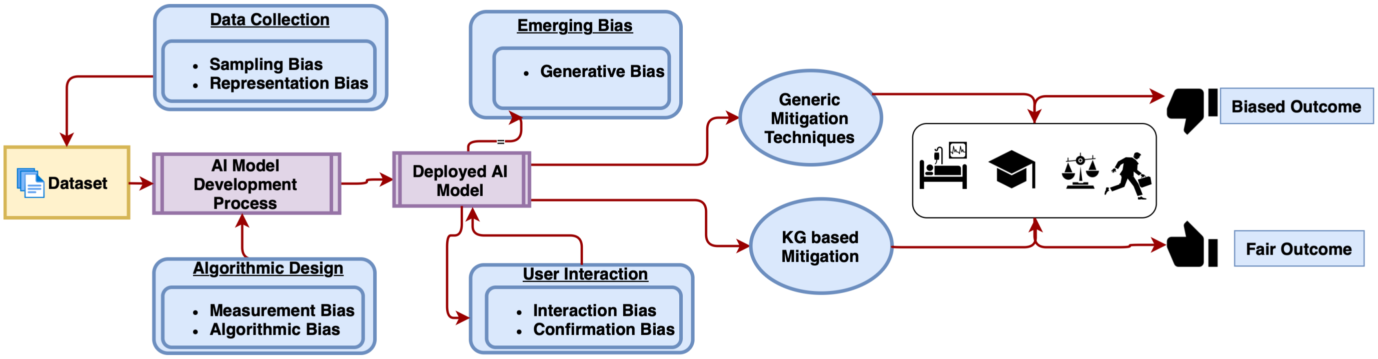


Fig. 4. Biases in the AI development and deployment lifecycle

(automation bias) reinforce existing systemic biases. Mansoury et al. [42] show in their research how the user interaction with recommendations cause a feedback loop that amplify existing biases in recommendation systems. Bias is also generated during the interaction process when the user tries to seek or generate information that aligns with their previously held beliefs, expectations, or assumptions which is known as confirmation bias. This can be caused via biased prompting, where users unintentionally guide the system toward responses that reinforce their views. Over time, this can create feedback loops that amplify existing biases [43].

d) Emerging Bias: Emerging bias refers to biases that surface when an AI model interacts with real-world users, often becoming apparent only after deployment. The most prominent type is generative bias.

Generative bias arises during the model output phase, where generative AI systems produce stereotypical, unfair, or harmful content. This bias reflects and often amplifies patterns from training data, manifesting in text, image, or audio generation. Zhou et al.[44] in their research reveal two key types of generative biases present in AI image generators: overt gender and racial biases, as well as more subtle biases related to facial expressions and physical appearance.

V. Impact of Biases

Bias in AI systems is more than just a technical flaw - it has real-world consequences that can negatively affect marginalized communities. From criminal justice to healthcare and online platforms, biased algorithms can reinforce existing inequalities and create new forms of discrimination.

COMPAS, a recdedivism tool used in the US, was found to be biased against African-American defendants. The study by [10] concluded that the COMPAS system used for recidivism prediction was only 66.07% accurate. The research found that African Americans were disproportionately assigned higher risk scores compared to Caucasians, despite being no more likely - and often less likely - to recidivate. This indicates a racial bias in a assessment tools.

Biases can lead to unfair outcomes for marginalized groups, where some individuals have an unfair advantage and some have an unfair disadvantage. Studies, such as the one done by [36], show that the datasets and algorithms used can perpetuate socio-economic biases. For instance, Uber's and Lyft's discriminatory dynamic pricing in locations where the majority of the population is African-American, longer waiting times for African-Americans, Facebook's STEM career ads that exclusively target men, and Optum's medical algorithms that are biased against Black customers based on their race [36]. Other examples of discriminatory outcomes are African-American customers experiencing higher cancellation rates, neighborhoods with lower socio-economic demographics being systematically excluded from Amazon's delivery system, and restricted access to critical resources or opportunities for certain socio-cultural groups due to perpetuated historical discriminatory practice.

A study by [45] revealed bias in an AI system designed to forecast patient mortality rates disproportionately affecting African-American patients. The system tended to assign higher-risk scores to African-American patients compared to others with similar age and health conditions. This biased scoring could result in African-American patients being overlooked for healthcare services or receiving inferior care, highlighting a significant racial bias in healthcare predictive analytics

LLMs are also a huge part of the AI landscape these days, and for most people that is how they interact with AI. A study by [33] addressed the fact that resources such as Wikipedia and OpenStreetMap get most of their data from the urban areas, with little contribution from the rural areas. Their objective was to analyze the geographic and socio-economic representation in LLMs and assess response consistency across regions, identifying areas of over or under-representation. The study found that LLM-generated responses tend to underrepresent historically marginalized groups, including non-White populations, individuals with disabilities, low-income or high-unemployment areas, and communities with limited financial resources. When prompted with location-based queries related to relocation, entrepreneurship, and tourism, the models consistently favored cities with higher economic status and lower demographic diversity. Additionally, the outputs exhibited a high degree of homogeneity in both content and language, suggesting shared biases in training data and model architecture that may contribute to cultural and economic monocultures.

Another study by [8] on the potential of LLM such as GPT-4 to perpetuate bias in clinical diagnosis showed that: the model significantly misrepresented disease prevalence across race and gender, often overrepresenting Black and female patients in conditions like Sarcoidosis and rheumatoid arthritis, while underrepresenting Hispanic and Asian populations except in stereotyped conditions such as hepatitis B and tuberculosis. The LLM also overrepresented female patients in diseases with female predominance but also demonstrated inconsistency in conditions with equal prevalence. Diagnostic rankings varied by gender, with certain conditions—such as anxiety—ranked higher for women regardless of clinical presentation. Bias was also seen in treatment recommendations and perceived patient honesty, with White males rated as more likely to exaggerate pain and Black males more likely to be flagged for drug misuse. These findings show existing disparities in LLM-generated clinical reasoning and potentially reinforcing healthcare inequities.

These examples show the need to address biases in AI driven systems. As biases can arise at any stage of the AI pipeline - from data collection and model design to deployment. Without active monitoring and intervention, automated systems can perpetuate harm under the guise of objectivity, further exacerbating social and economic disparities.

VI. Handling Biases at each step of the AI Lifecycle

Taking a lifecycle-based approach to bias mitigation clarifies how biases arise at each phase of AI development—be it in data collection, model training, or user interaction—and guides mitigation strategies that tackle them at their source. This helps to target the bias at their origin in the AI lifecycle.

a) Data-Centric Mitigation (Pre-processing): Data-centric methods address bias before the model is trained. These strategies typically involve data balancing (e.g., oversampling underrepresented groups), relabeling, or perturbation [5, 13]. For example, oversampling darker-skinned individuals in facial recognition training sets has improved model accuracy for those groups [46] . However, modifying datasets can be time-consuming and may pose legal or ethical challenges when dealing with sensitive personal data.

b) Model-Centric Mitigation (In-processing): Model-centric methods aim to limit bias during training. Techniques include adversarial learning, regularization with fairness constraints, and compositional approaches that reduce a model's reliance on sensitive attributes [35]. In a survey of 314 publications, in-processing methods were found to be the most common bias mitigation approach—appearing twice as frequently as pre-processing methods and four times as often as post-processing [13]. This prevalence may be due to ease of integration within existing training pipelines and direct control over model behavior.

c) Output-Centric Mitigation (Post-processing): Post-processing adjustments modify or recalibrate model outputs to meet fairness requirements, such as Equalized Odds or Equal Opportunity [5, 35]. Despite being least utilized in published work [13], post-processing can be crucial when retraining a model from scratch is infeasible or when quick fixes are needed for existing systems. However, certain studies caution that enforcing specific fairness metrics may produce counterintuitive results if not carefully optimized [47].

VII. Trends in Bias Research

This section presents a comprehensive analysis of global trends in AI research, highlighting overarching patterns such as co-authorship networks and the distribution of research efforts across regions. The analysis is structured around five categorized thematic areas: (1) *Health & Clinical AI*; (2) *Recommender Systems*; (3) *LLMs & NLP*; (4) *General Fairness & Bias Mitigation*; and (5) *Graph-Based Fairness & Bias Mitigation*. This study makes a distinctive contribution by highlighting regional disparities and thematic gaps within these research areas, thereby offering valuable insights into the global landscape of AI bias research. Our insights are drawn from a comprehensive review of 99 research papers included in this survey. To support deeper exploration and transparency, we also provide an interactive dashboard that allows researchers to explore the data and findings in detail. The interactive dashboard and its source code are available at [48] and [49], respectively.

A. Distribution of Research Across Thematic Domains

Fig. 5 shows the thematic distribution of our categorized research areas: (1) *Health & Clinical AI*; (2) *Recommender Systems*; (3) *LLMs & NLP*; (4) *General Fairness & Bias Mitigation*; and (5) *Graph-Based Fairness & Bias Mitigation*. As shown in Fig. 5, the majority of research is dedicated to general fairness and bias mitigation strategies—such as debiasing word-level embeddings—accounting for 47.5% (47 out of 99 papers). This highlights a strong emphasis on addressing bias in task-agnostic contexts and at the foundational level, such as within natural language processing algorithms. Following this, the *LLMs & NLP* category accounts for 22.2% (22 out of 99 papers), addressing biases emerging from NLP algorithms and large language models. Examples include issues such as LLM hallucination and social bias in sentence-level representations, as examined by Liang et al.[50]. Healthcare and Clinical AI represents the third largest category, comprising 15.2% (15 out of 99 papers), and focuses on research addressing the detection and mitigation of racial bias in AI-powered systems deployed in healthcare. The fourth category, *Recommender Systems*, accounts for 7.1% (7 out of 99 papers) and examines biases arising within recommender systems, such as disparities in how active and inactive users are treated, along with strategies for their mitigation. The *Graph-Based Fairness & Bias Mitigation* category, which includes approaches leveraging semantic web technologies—for example, the method proposed by [17]—accounts for only 8.1% (8 out of 99 papers).

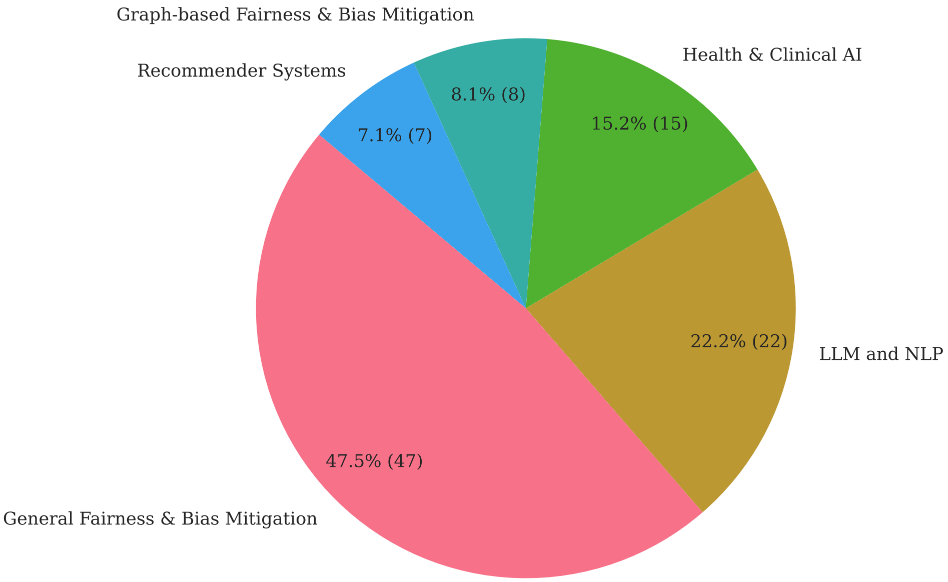


Fig. 5. Distribution of research based on categorized domains: (1) *Health & Clinical AI*; (2) *Recommender Systems*; (3) *LLMs & NLP*; (4) *General Fairness & Bias Mitigation*; and (5) *Graph-Based Fairness & Bias Mitigation*

B. Temporal Dynamics of Bias Mitigation Research Across Domains

Fig. 6 and Fig. 7 provide detailed insights into the evolution of research on AI bias over time. An interactive version of these figures is available through the accompanying dashboard [48]. Specifically, Fig. 6 illustrates the overall temporal progression of publications, whereas Fig. 7 shows this progression by research domain, offering a domain-specific perspective on trends over time.

As shown in Fig. 6, both the number of publications and the diversity of categorized thematic domains grew gradually after 2015, with a marked acceleration beginning in 2019 and culminating in a peak in 2023. For instance, in 2019 there were 5 publications spanning 2 domains, a figure that more than doubled the following year to 13 publications across 4 domains. A similar upward trajectory is observed in subsequent years, with the number of publications steadily increasing and the coverage of thematic domains fluctuating. By 2023, research output reached its highest point with 21 publications encompassing all five categorized domains. The year 2024, however, saw the drop in the number of publications.

A comparable trend is observed in the domain-wise progression of research, as shown in Fig. 7, where each domain exhibits a growth trajectory broadly aligned with the overall trend of AI bias research (see Fig. 6). However, closer observation reveals differences in how individual domains have evolved over time. The domain *General Fairness & Bias Mitigation* demonstrates the most consistent growth in publication volume, peaking in 2021 with 16 publications before declining in subsequent years. In contrast, *LLMs & NLP* show a sharp increase, reaching their peak in 2023 with 10 publications, a rise that can be attributed to the development of LLMs and their applications. Similar to *General Fairness & Bias Mitigation*, this domain also exhibits a decline following its peak. The *Health & Clinical AI* domain presents a fluctuating trajectory, with gradual growth between 2018 and 2020, a sharp decline in 2021, a peak of 6 publications in 2022, and a subsequent decrease. The *Recommender Systems* domain, by comparison, maintains a relatively steady pattern with minimal variation across years. Finally, the *Graph-Based Fairness & Bias Mitigation* domain has demonstrated an upward trend since 2020.

C. Institutional Contributions Within Research Domains

Fig. 8 and Fig. 9 present the top 20 institutional contributions across the thematic domains of AI bias research. Fig. 8 presents contributions based on all authors, whereas Fig. 9 restricts the analysis to first-author contributions. We include all authors, not just the first author, in our analysis to capture the overall scope of participation in research. This approach allows us to assess the institutional emphasis on specific research areas, revealing both the size of involvement and the domains in which particular institutions prioritize their efforts.

As shown in Fig. 8, University College London and the Rutgers University emerge as the leading institutions in the domain of *General Fairness & Bias Mitigation*, contributing 25.7% and 12.2%, respectively. North Carolina State University follows with 9.5%, while Michigan State University and the University of South California each contribute 8.1%. An interesting observation arises within the same institution across different geographic regions: for instance, IBM Research contributes 4.1%, whereas IBM Research India accounts for 6.8%. Among the top 20 institutions, MIT, Stanford University,

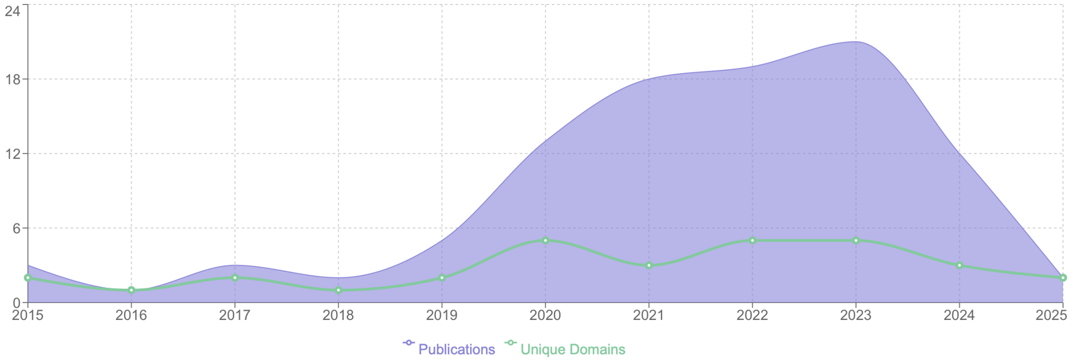


Fig. 6. Temporal Progression of Scholarly Publications and Domain Coverage in AI Bias Research (2015–2025)

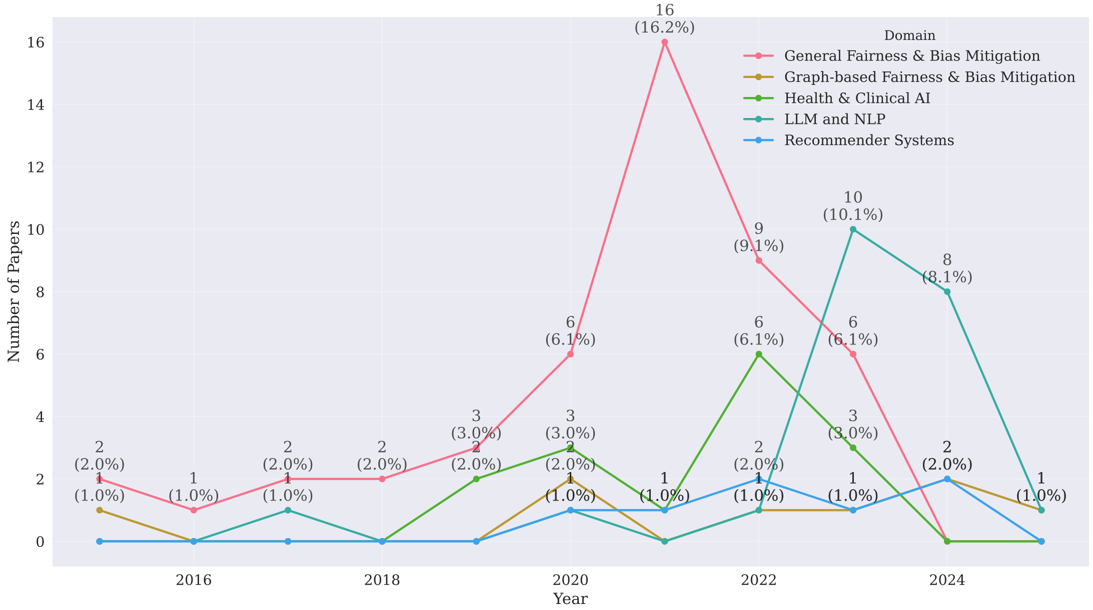


Fig. 7. Shifting Emphasis in Bias Mitigation Research Across Domains Over Time.

Harvard University, The Hong Kong Polytechnic University, and Tsinghua University represent the lowest contributors in this domain, each with a share of 1.4%. In the domain of *Health & Clinical AI*, Harvard University leads with 42.9%, followed by MIT with 28.6%. University of Cagliari, University of Pennsylvanian, Adobe, and Carnegie Mellon University each contribute 7.1%. These results highlight distinct institutional priorities in bias research. Notably, Harvard University—one of the lowest contributors in *General Fairness & Bias Mitigation*—emerges as the top institution in *Health & Clinical AI*, reflecting a strong concentration of researchers in this area. In the case of the *LLMs & NLP* domain, Tsinghua University emerges as the leader with 22.7%. Northeastern University, Carnegie Mellon University, and Adobe each contribute 13.6%. Similarly, UCLA, IBM Research, and the University of Hong Kong each account for 6.8%, while University College London and the University of Washington contribute 4.5% each. The lowest share, 2.3%, is held by Harvard University, MIT, and the University of Southern California. For the *Recommender Systems* domain, the Rutgers University leads with 63.8%, followed by the University of Cagliari and The Hong Kong Polytechnic University, each contributing 10.6%. The University of Hong Kong accounts for 6.4%, while Michigan State University contributes 4.3%. The lowest shares, 2.1% each, are held by UCLA and the University of Washington. In the *Graph-based Fairness & Bias Mitigation* domain we see that Leibniz University Hannover solely dominates with 100% contribution.

In the case of first-author contributions, as shown in Fig. 9, a distinct pattern emerges compared to Fig. 8, highlighting differences in research concentration when considering only the first author versus all contributing authors.

For *General Fairness & Bias Mitigation*, University College London leads with 31.6%, consistent with the results for all authors. The University of Pennsylvania, the University of Southern California, Iowa State University, and North Carolina State University share the second position, each contributing 10.5%. Notably, Iowa State University appears here despite not ranking among the top institutions when considering all authors. This indicates that, although some institutions may have a smaller number of researchers, they nonetheless play a significant role in leading the work and contribute to the

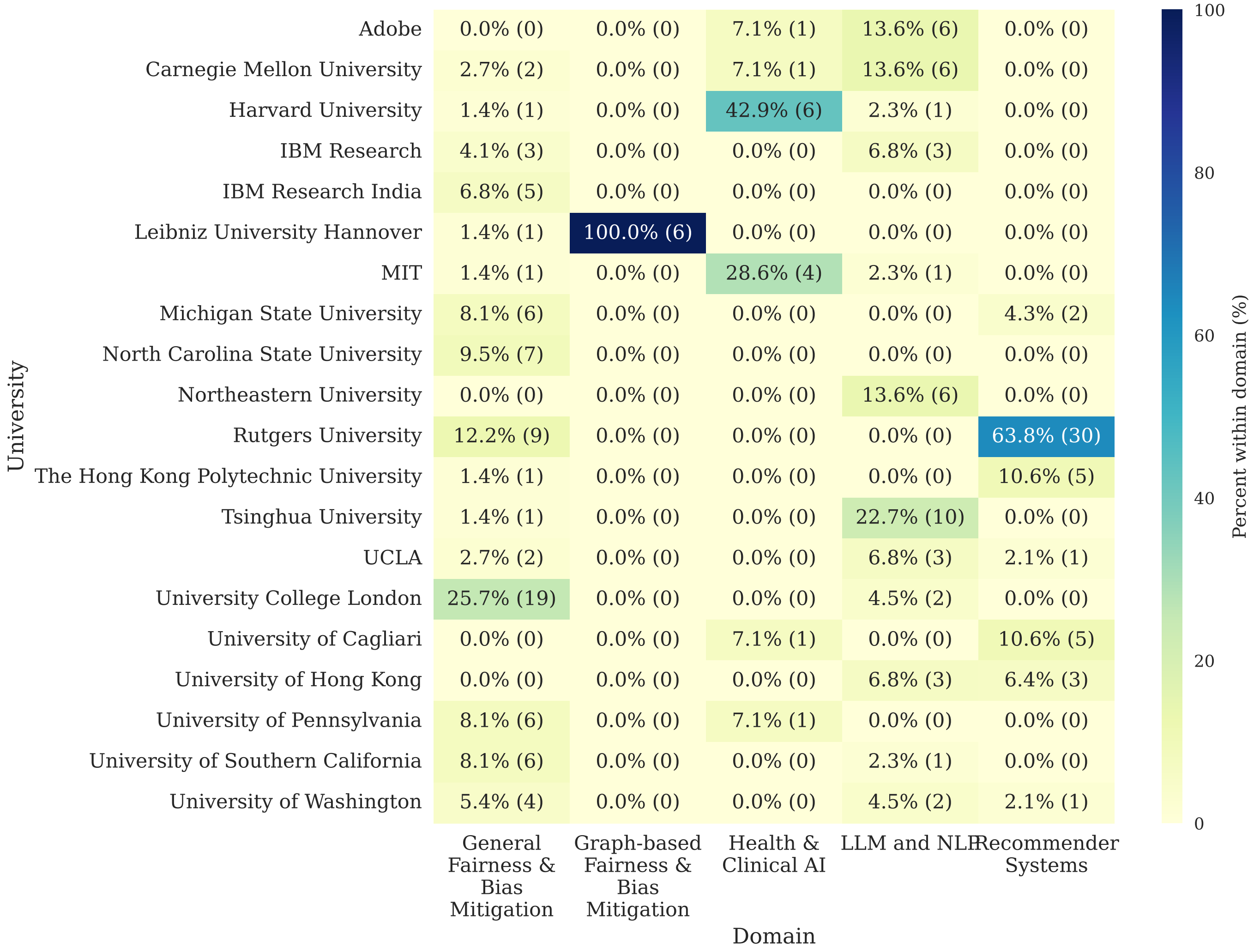


Fig. 8. Top 20 institutional contributions to AI bias research across domains, aggregated over all authors. Each paper with multiple authors from the same university is counted once per author, such that institutions with multiple co-authors receive multiple counts. Columns represent the share of author–affiliation assignments within each domain. Cell labels are reported as "% (n)", where n denotes the number of author–affiliation assignments.

the field. Similarly, we observe new institutions such as the Academic College of Tel Aviv–Yafo, Rutgers, and Cornell University, each with 5.3%, alongside MIT, Carnegie Mellon University, Stanford University, and University of Pennsylvania. In the case of *Graph-based Fairness & Bias Mitigation* Althire AI, Columbia University, and Deakin University each make an equal contribution of 33.3%. Here, Leibniz University is not seen, indicating that although the university had the most contributions while considering all authors, when only first authors are considered other institutions take the lead. Similarly, in the case of *Health & Clinical AI*, Harvard University no longer leads the ranking when only first authors are considered. Instead MIT, the University of Pennsylvania, the Fraunhofer Institute for Digital Medicine, Carnegie Mellon University, the Cardiovascular Institute of the South, Cape Breton University, Brigham and Women's Hospital, and Bennett University have share an equal contribution of 12.5%. This suggests that Harvard University's prominence in this domain is driven by a larger pool of contributing researchers rather than first-author leadership. In the case of the *LLMs & NLP* domain, Stanford University emerges as the leader with 50%, followed by Carnegie Mellon University and Florida International University, each contributing 25%. For *Recommender Systems*, the University of Rutgers University dominates with 100%. Similar to the *Health & Clinical AI* domain, these results highlight a notable shift in institutional research focus, particularly the concentration of efforts at Stanford University and the University College London.

D. Domain Participation of Leading Countries

Fig. 10 and Fig. 11 present the domain-specific, country-wise collaborations in AI bias research, based on all authors and first authors, respectively.

As shown in Fig. 10, within the *General Fairness & Bias Mitigation* domain, the United States leads with a high researcher concentration of 42.2%, followed by the United Kingdom at 17.2% and Germany at 11.1%. Italy accounts for 3.9%, followed by France (3.3%) and India (2.8%). Israel, South Korea, Greece, and Australia each contribute 2.2%,

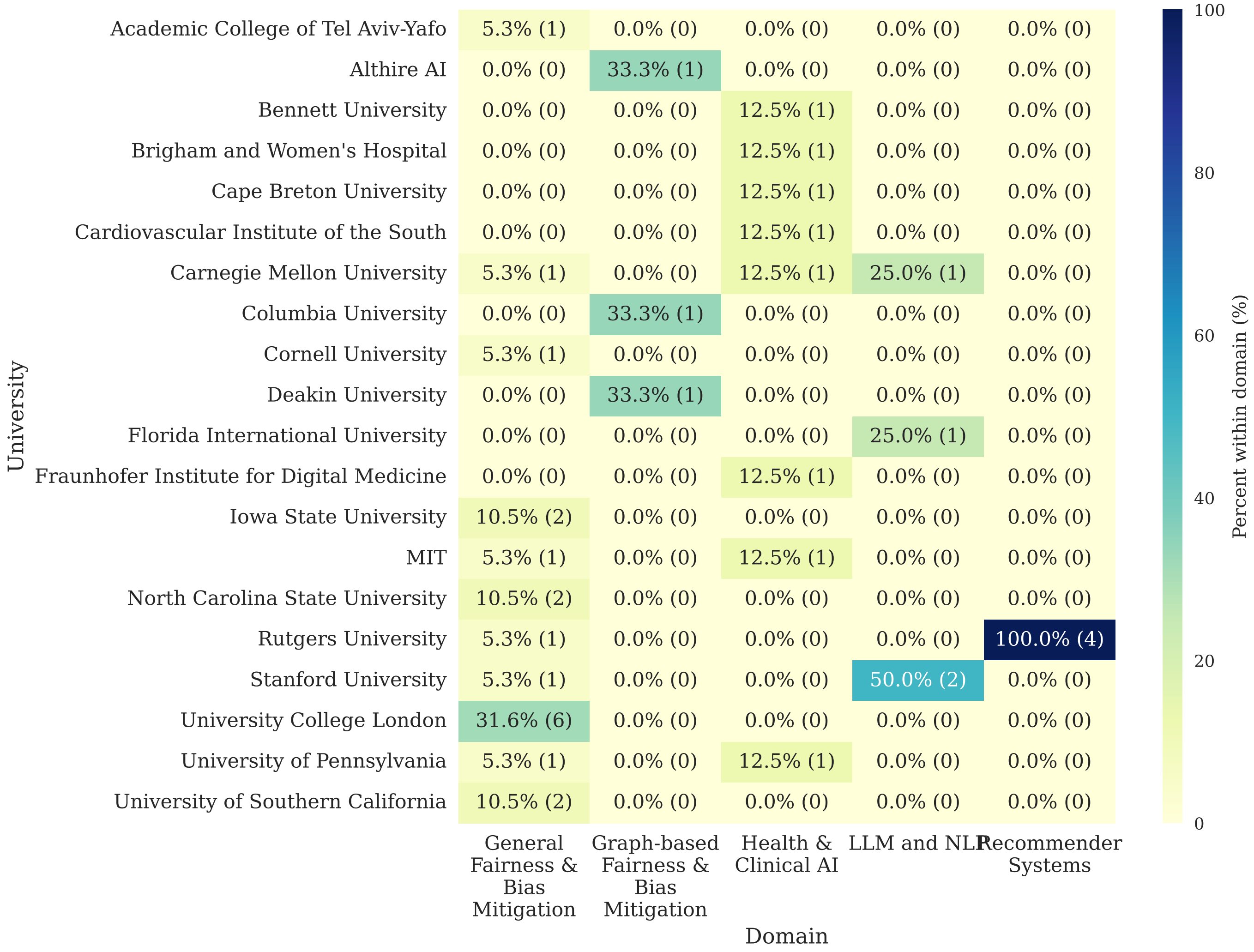


Fig. 9. Top 20 institutional contributions to AI bias research across domains, based on first-author affiliations only. Each paper is counted once, according to the affiliation of its first author. Columns represent the share of first-author affiliations within each domain. Cell labels are reported as "% (n)", where n denotes the number of first-author assignments.

while the remaining countries such as Austria, Canada, and the Netherlands have minimal involvement (less than 2%). In the *Graph-Based Fairness & Bias Mitigation* domain, the distribution is concentrated, with the Germany accounting for 40.9% followerd by the UK for 18.2% and the USA and Greece for 11.4% and 9.1% respectively. For *Health & Clinical AI*, the United States again takes the lead with 45.3%, followed by Germany (18.9%). Switzerland and India each hold 6.6%, while Austria accounts for 3.8%, and Canada and Finland share 2.8% each. The United Kingdom, the Netherlands, Australia, and France each contribute 1.9%, reflecting broader but lower-level international engagement. Italy and Denmark make the lowest contribution with 0.9% each.

In *LLMs & NLP*, consistent with trends in other domains, the United States dominates with 48.6%, followed by China with 33.3%. Other contributors include Canada (3.5%), Spain and Taiwan (each 2.8%), Israel and Denmark (each 2.1%), and Hong Kong (1.4%). Finally, in the *Recommender Systems* domain, the United States again ranks first with 64.2%, followed by Hong Kong (17%) and Italy (9.4%). The United Kingdom contributes 5.7%, and Canada accounts for 3.8%, highlighting shifting national priorities across domains of AI bias research.

In the case of first-author only contributions (see Fig. 11), a broadly similar pattern to that observed for all authors emerges, particularly with respect to the countries leading research within each domain. The United States consistently leads across all domains. The United Kingdom follows in second place in the *General Fairness & Bias Mitigation* domain and shares an equal proportion of 50% with France in *Graph-Based Fairness & Bias Mitigation*. For *Health & Clinical AI*, Switzerland, India, Australia, Canada, and Germany each contribute 6.7%, indicating equal levels of participation when considering only first authorship. In *LLMs & NLP*, the pattern remains consistent, with USA dominating at 50.0%, China occupying the second position at 22.7% after the United States. Canada, Denmark, Israel, Spain, Taiwan, and the United Kingdom each contribute 4.5%. In the case of *Recommender Systems*, contributions are more evenly distributed after the United States, with the United Kingdom, Italy, and Hong Kong each accounting for 14.3%.

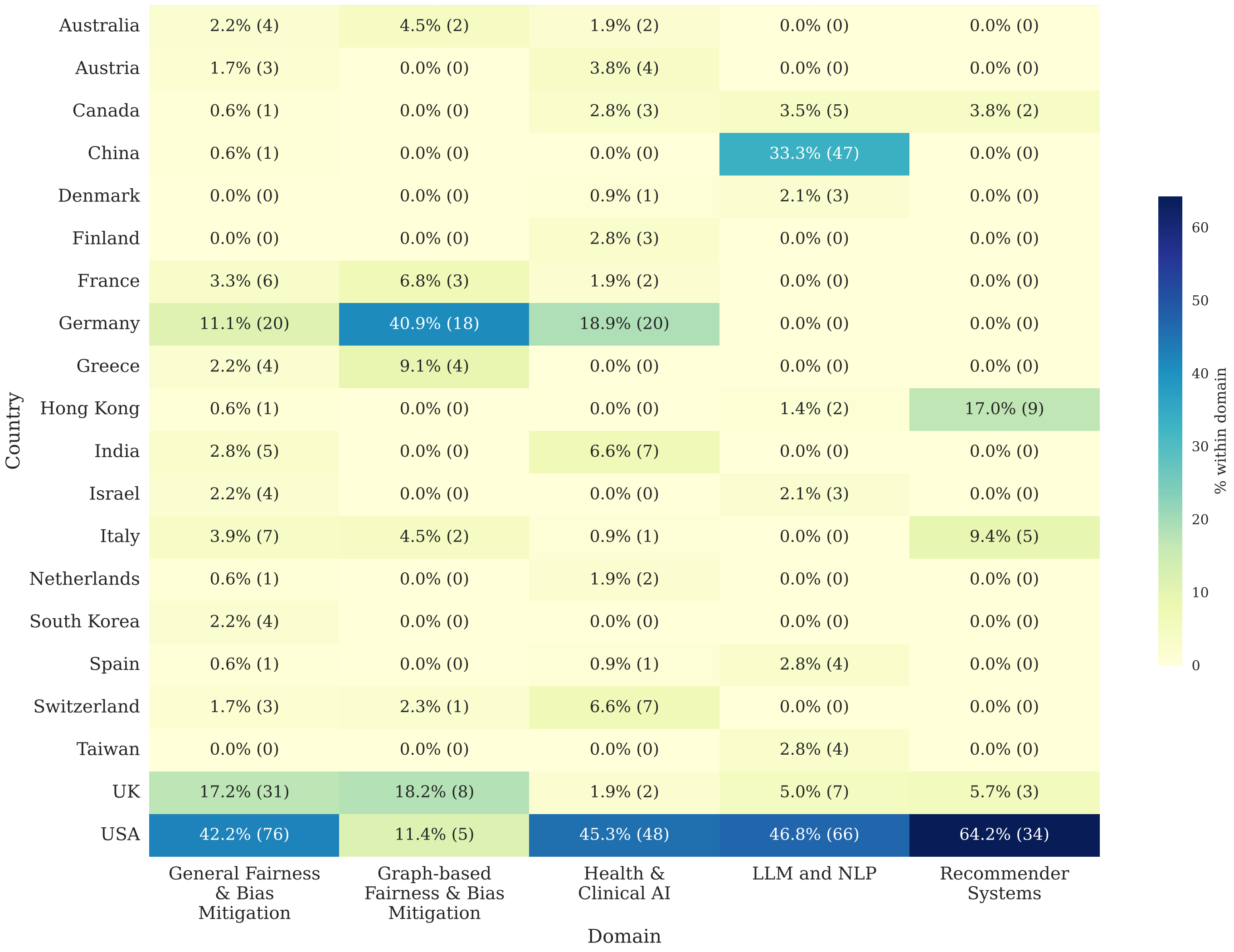


Fig. 10. Top 20 country-wise collaborations in AI bias research across domains (all authors). Each paper with multiple authors from the same country is counted once per author, such that countries with multiple co-authors receive multiple counts. Columns represent the share of author–country assignments within each domain. Cell labels are reported as "% (n)", where n denotes the number of author–country assignments.

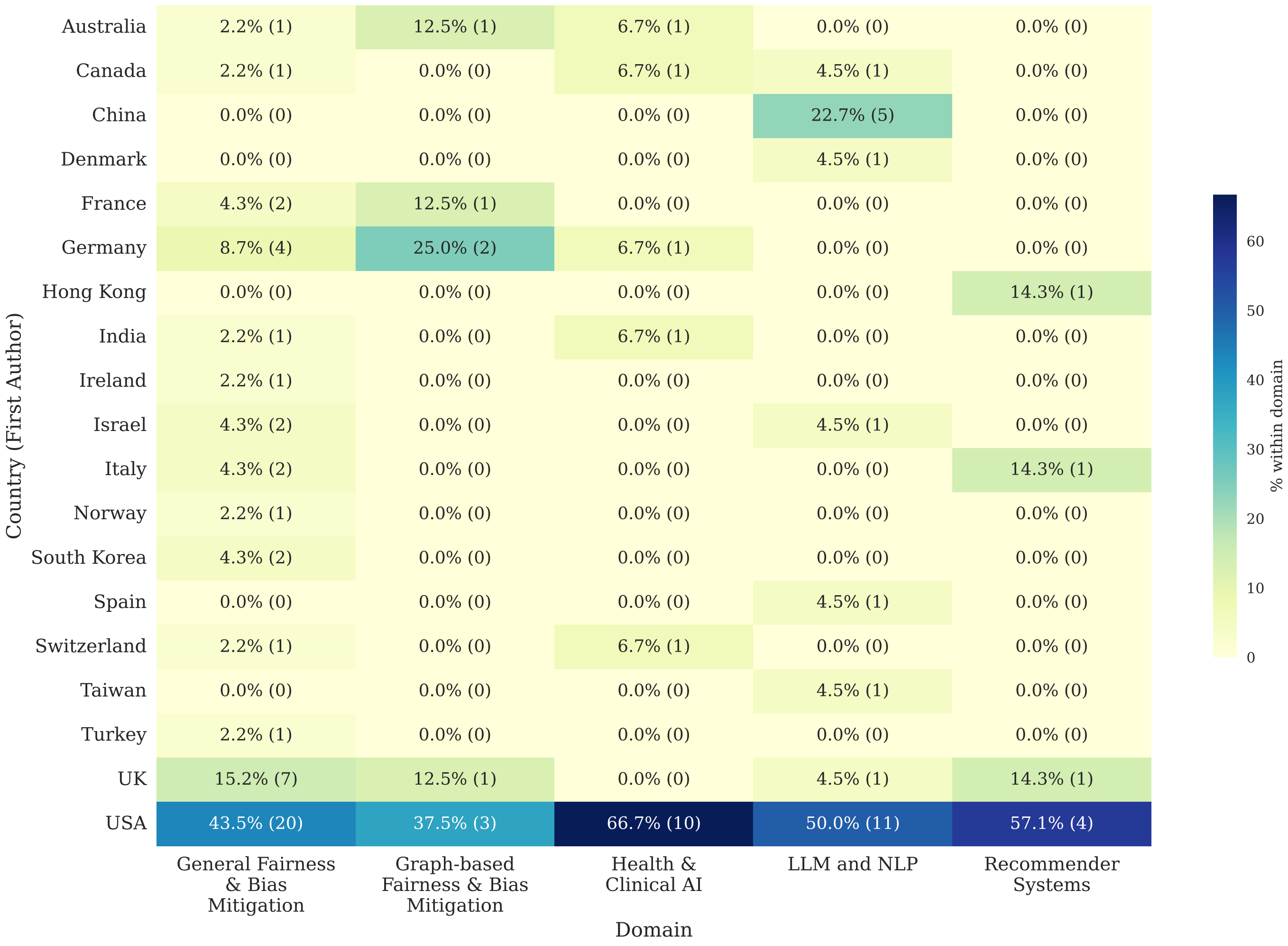
E. International Collaboration Networks

Fig. 12 presents the heatmap of cross-country collaboration. The figure illustrates that the United Kingdom (26 collaborations with 18 partner countries) and the United States (23 collaborations with 12 partner countries) emerge as the principal hubs within the collaboration network. Their extensive and dense connections highlight their central roles in facilitating cross-national research partnerships. Similarly, Austria, the Netherlands, Germany, and Italy form a dense European cluster, underscoring their roles as core actors within the region. Among them, Austria emerges as a central hub with 19 collaborations across 16 countries, followed by the Netherlands with 14 collaborations across 12 countries, and Germany with 17 collaborations across 14 countries. Italy also demonstrates significant engagement, recording 14 collaborations across 10 countries. Greece has 10 collaborations with 8 countries. At the next tier, Belgium has 7 collaborations with 7 countries, while Denmark, Finland, and Spain follow with 6 collaborations across 6 countries. Sweden, Poland, and Switzerland maintain 5 collaborations with 5 partners, and France records 4 collaborations across 4 countries. At the lower end of the spectrum, Ireland, Norway, and Israel are represented by a single collaboration each.

In Asia, China holds the leading position with 10 collaborations across 3 countries, followed by Japan with 6 collaborations across 6 countries, Hong Kong with 5 collaborations across 2 countries, and the United Arab Emirates (UAE) with 5 collaborations across 5 countries. India occupies the lowest position in the region, with 3 collaborations in 3 countries. Beyond Asia, Australia and Canada each match China in terms of total collaborations (9), though they differ in breadth: Australia engages with 8 partner countries, whereas Canada collaborates with 7. Chile also demonstrates notable participation, recording 7 collaborations across 7 countries.

F. Country-Level Authorship and Publication Patterns

Fig. 13 and Fig. 14 provide an overview of participation in AI bias research by country, measured in terms of author



Country (First Author)	General Fairness & Bias Mitigation	Graph-based Fairness & Bias Mitigation	Health & Clinical AI	LLM and NLP	Recommender Systems
Australia	2.2% (1)	12.5% (1)	6.7% (1)	0.0% (0)	0.0% (0)
Canada	2.2% (1)	0.0% (0)	6.7% (1)	4.5% (1)	0.0% (0)
China	0.0% (0)	0.0% (0)	0.0% (0)	22.7% (5)	0.0% (0)
Denmark	0.0% (0)	0.0% (0)	0.0% (0)	4.5% (1)	0.0% (0)
France	4.3% (2)	12.5% (1)	0.0% (0)	0.0% (0)	0.0% (0)
Germany	8.7% (4)	25.0% (2)	6.7% (1)	0.0% (0)	0.0% (0)
Hong Kong	0.0% (0)	0.0% (0)	0.0% (0)	0.0% (0)	14.3% (1)
India	2.2% (1)	0.0% (0)	6.7% (1)	0.0% (0)	0.0% (0)
Ireland	2.2% (1)	0.0% (0)	0.0% (0)	0.0% (0)	0.0% (0)
Israel	4.3% (2)	0.0% (0)	0.0% (0)	4.5% (1)	0.0% (0)
Italy	4.3% (2)	0.0% (0)	0.0% (0)	0.0% (0)	14.3% (1)
Norway	2.2% (1)	0.0% (0)	0.0% (0)	0.0% (0)	0.0% (0)
South Korea	4.3% (2)	0.0% (0)	0.0% (0)	0.0% (0)	0.0% (0)
Spain	0.0% (0)	0.0% (0)	0.0% (0)	4.5% (1)	0.0% (0)
Switzerland	2.2% (1)	0.0% (0)	6.7% (1)	0.0% (0)	0.0% (0)
Taiwan	0.0% (0)	0.0% (0)	0.0% (0)	4.5% (1)	0.0% (0)
Turkey	2.2% (1)	0.0% (0)	0.0% (0)	0.0% (0)	0.0% (0)
UK	15.2% (7)	12.5% (1)	0.0% (0)	4.5% (1)	14.3% (1)
USA	43.5% (20)	37.5% (3)	66.7% (10)	50.0% (11)	57.1% (4)

Fig. 11. Top 20 country-wise contributions to AI bias research across domains (first author only). Columns represent the share of papers within each domain whose first author is affiliated with each country. Cell labels are reported as "% (n)", where n denotes the number of first-authored papers for that country–domain pair.

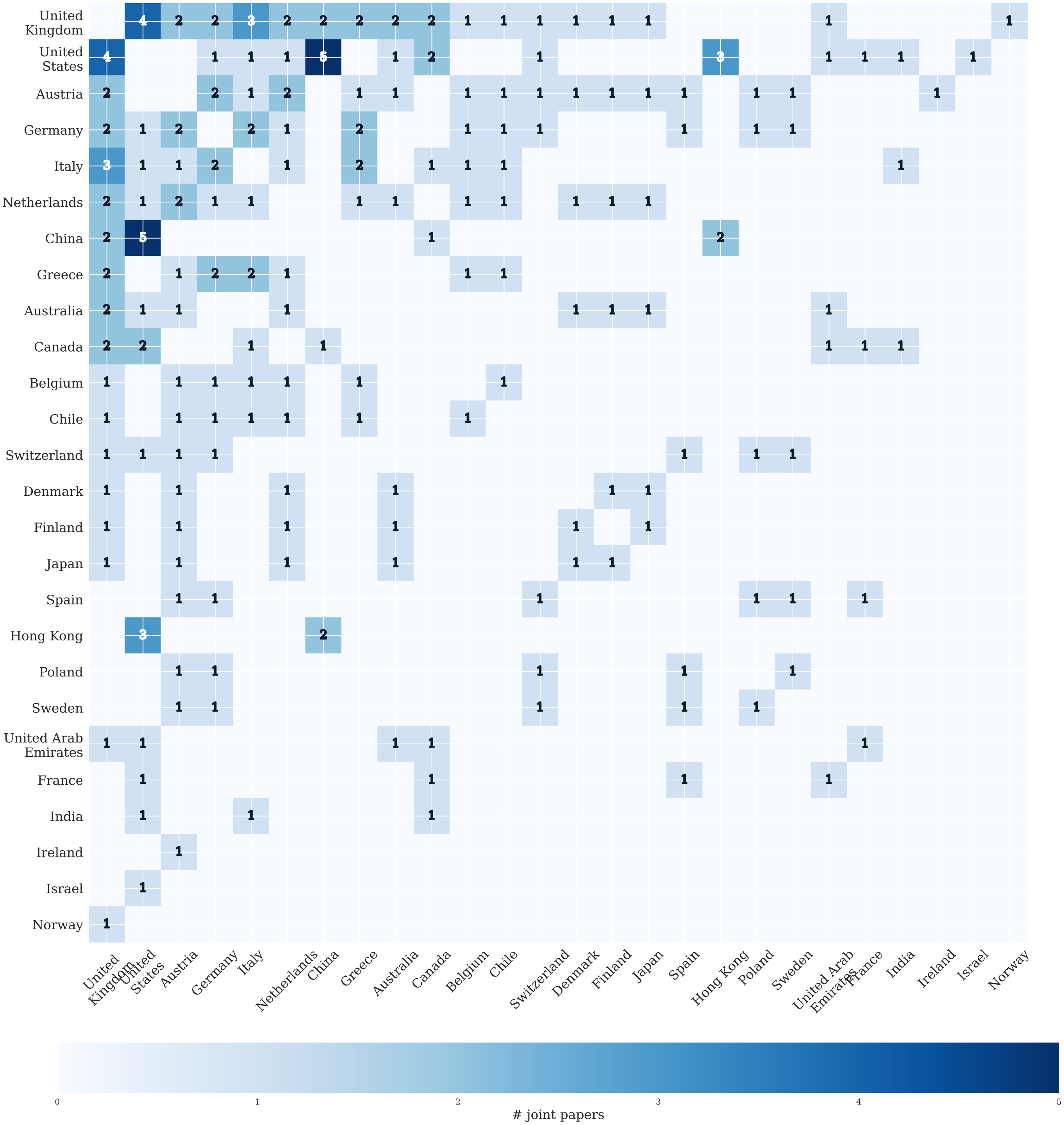
counts and publications. Fig. 13 includes all authors, while Fig. 14 focuses exclusively on first authors.

Consistent with earlier analyses, the United States dominates, with 228 authors contributing to 56 papers. The United Kingdom follows with 51 authors across 19 publications, and China with 48 authors contributing to 8 publications. Canada ranks next with 6 publications, while Germany records 7 publications supported by 56 authors—indicating extensive multi-author involvement, with an author-to-paper ratio as high as 8. Similarly, Italy (15 authors, 6 publications), France (11 authors, 5 publications), Greece (8 authors, 2 publications) Austria (7 authors, 4 publications), Hong Kong (12 authors, 3 publications), Switzerland (10 authors, 3 publications), Israel (7 authors, 3 publications), Spain (6 authors, 3 publications), India (12 authors, 2 publications), and Australia (8 authors, 3 publications) also demonstrate notable participation. For other countries, contributions are comparatively smaller, typically ranging from 4 authors with 2 publications, 4 with 1, 3 with 1, or single-author single-publication contributions.

In the case of the first-author only analysis (see Fig. 14), a similar pattern to that observed for all authors is evident. The United States continues to lead, with 44 first authors contributing to 48 publications. The United Kingdom follows with 7 first authors across 10 publications, while Germany records 6 first authors leading 7 publications and China with 5 first authors and 5 publications. Other countries, including Australia, Canada, France, and Italy, also show balanced contributions with 3 first authors each leading 3 publications, except for Israel, where 2 first authors account for 3 publications. India, South Korea, and Switzerland each report a 2:2 author-to-paper proportion, while the remaining countries are represented by single-author, single-publication contributions.

VIII. Use of Ontologies and KGs in Bias Mitigation

Lobo et al. [17] present various scenarios where ontologies have been utilized to deal with bias in AI systems. One of the presented solutions is to leverage KGs to extract patterns from the graph, e.g., to capture spurious model correlations that are based on sensitive information, or properties that enable mining less popular items in a recommender system. Applying this concept to healthcare diagnostics, KGs can help uncover rare



	United Kingdom	United States	Austria	Germany	Italy	Netherlands	China	Greece	Australia	Canada	Belgium	Chile	Switzerland	Denmark	Finland	Japan	Spain	Hong Kong	Poland	Sweden	United Arab Emirates	France	India	Ireland	Israel	Norway
United Kingdom		4	2	2	3	2	2	2	2	2	1	1	1	1	1	1					1					1
United States	4			1	1	1	5		1	2			1					3			1	1	1		1	
Austria	2			2	1	2		1	1		1	1	1	1	1	1	1		1	1				1		
Germany	2	1	2		2	1		2			1	1	1				1		1	1						
Italy	3	1	1	2		1		2		1	1	1											1			
Netherlands	2	1	2	1	1			1	1		1	1		1	1	1										
China	2	5								1								2								
Greece	2		1	2	2	1					1	1														
Australia	2	1	1			1								1	1	1					1					
Canada	2	2			1		1														1	1	1			
Belgium	1		1	1	1	1		1				1														
Chile	1		1	1	1	1		1			1															
Switzerland	1	1	1	1													1		1	1						
Denmark	1		1			1			1						1	1										
Finland	1		1			1			1					1		1										
Japan	1		1			1			1					1	1											
Spain			1	1									1						1	1		1				
Hong Kong		3					2																			
Poland			1	1									1				1			1						
Sweden			1	1									1				1		1							
United Arab Emirates	1	1							1	1												1				
France		1								1							1				1					
India		1			1					1																
Ireland			1																							
Israel		1																								
Norway	1																									

Fig. 12. Cross-country collaboration, calculated without author multiplicity (i.e., one collaboration counted per paper per country pair).

disease correlations, avoid spurious correlations, like biased patterns that depend on race or gender, with which healthcare providers can perform faster, more accurate diagnoses and facilitate personalized treatment based on individual genetics and lifestyle.

The Multilingual Visual Sentiment Ontology (MVSO) is another innovative approach to visual sentiment analysis by incorporating multilingual contexts into its framework [51]. This ontology captures cultural and linguistic variations in sentiment by addressing significant challenges in sentiment analysis. MVSO can help healthcare providers, particularly valuable in mental health services, better understand patient emotions from multilingual social media posts, patient feedback, and digital communication channels without being biased, or depending on heuristics.

In recent years, researchers have started using KGs to better understand and reduce bias in AI systems. KGs are powerful as they can capture connections and context and make the training

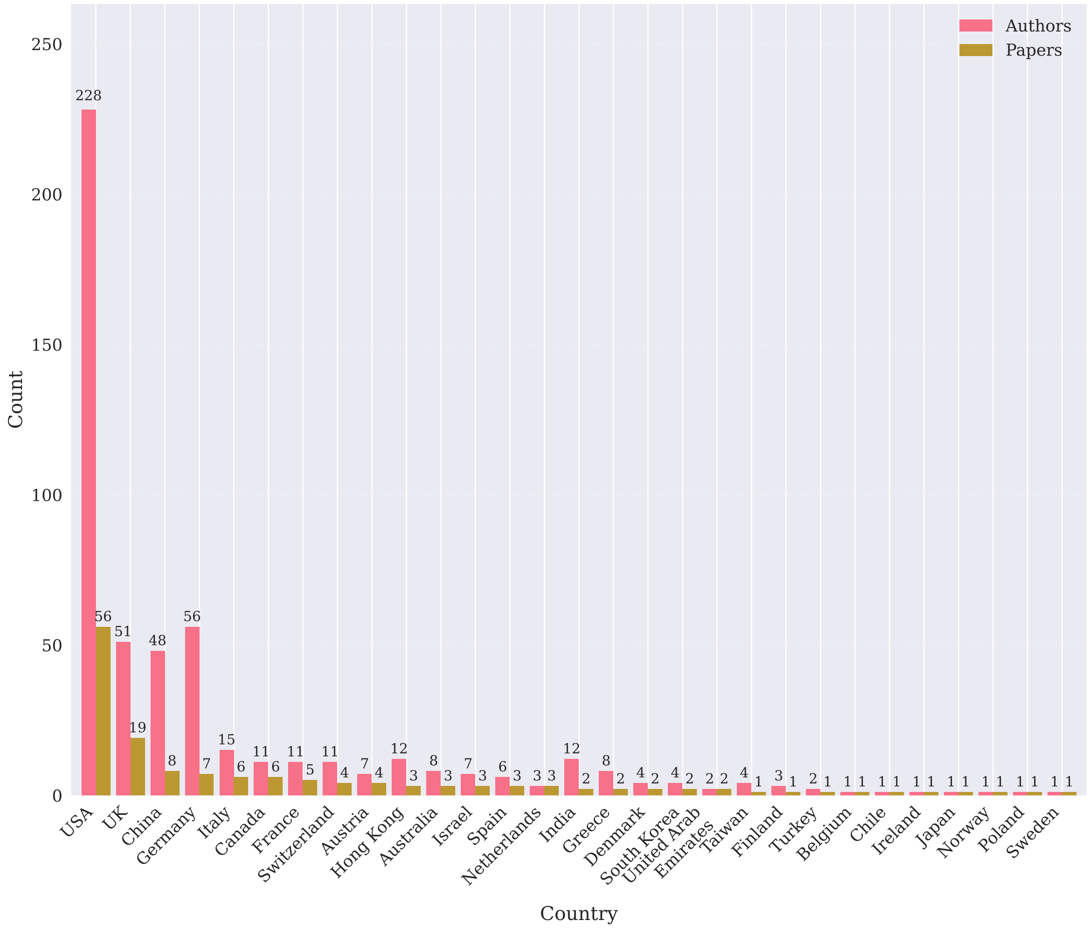


Fig. 13. Countrywise: number of total authors and papers

data more richer and context aware which might otherwise be missing. KGs can be used to create relation - social, semantic, or hierarchical, among different elements.Their proper application can make AI models more transparent and even help them make fairer decisions. This subsection explores how KGs are being applied to reduce bias across domains - from contextual data pre-processing to fairness-aware training of language models.

Huang et al. [20] present a novel use of contextual KGs (CKGs) to tackle bias in AI-based decision support systems. Their work addresses a critical pain point in fairness research: lack of contextual awareness in traditional ML pipelines. Basically, their experiment focused on debiasing the dataset itself, using a contextual knowledge graph (CKG) to guide the identification and correction of bias before training the model.

Kumar et al. [21] proposed a strong demonstration of how ontology-aware architectures can enhance both fairness and performance. In the experiment, the researchers have attempted to detect and mitigate biases in LLMs through KG-augmented training. They have addressed the issue of LLMs inheriting and even amplifying biases that are in the training data. They call it Knowledge Graph Augmented Training (KGAT). They ran their experiment on 3 datasets: a. The Bias in Bios dataset [52] for gender stereotype analysis, b. The CelebA dataset [53] for facial attribute classification, and c. The ProPublica COMPAS dataset [54] for fairness in recidivism prediction. The researchers encoded KGs into vector representations using Graph Neural Networks (GNNs), which were then integrated with LLMs using multi-head attention mechanisms. The results they produced were sufficient to validate their hypothesis that KGAT could effectively reduce biases while also improving model performance. Integrating KG allowed the models to make use of the relational knowledge which resulted in more equitable and unbiased predictions.

Deshpande et al. [55], have made use of a knowledge graph to address cultural stereotyping, which is a specific form of social bias. They created a knowledge graph named StereoKG, of cultural knowledge and stereotypes which covered 5 religious groups and 5 nationalities. Their research showed that performing an intermediate masked language model training on the verbalized KG lead to a more culturally aware model and has the potential to increase classification performance. This KG gives the model a structured exposure to what stereotypes look like, and as a result they perform significantly better at stereotype-sensitive tasks than baseline models. Unlike typical demographic fairness studies (e.g., gender or race in hiring), it captures free-form, crowd-sourced cultural knowledge from social media, which makes it context-rich and organic in terms of the cultural diaspora.

Gaur et al. [56] present a study in which the researchers used knowledge graphs to the interpretability and explainability

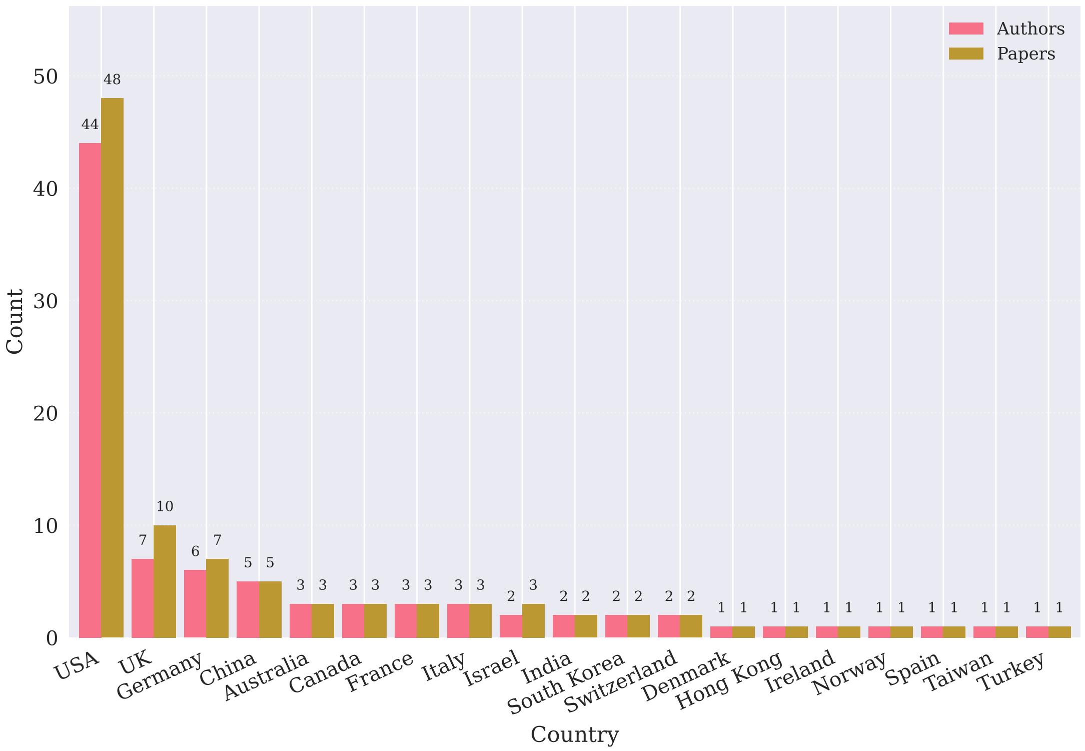


Fig. 14. Countrywise: number of total first authors and papers

of deep learning systems. The researchers emphasize that explainability is significant for highlighting inherent biases in predictive models and also prevent prediction errors in unintuitive scenarios. Their framework was applied in the domain of education to address the challenge of predicting student performance, important for evaluating true potential. Infusing a structured knowledge base, either in a shallow or a deep manner as stated above, can help trace a decision back to its roots cause increasing explainability and accountability. This can directly help in debiasing biased outcomes, for example: a biased output could be traced back and updates can be made to the knowledge base to make the output fair.

In summary, KG-based approaches have so far shown promising results for improving fairness and mitigating biases in AI systems by embedding contextual information, semantics, and structure. Most of the KG-ontology based mitigation methods show that the accuracy of the system was improved along with the fairness. Table II presents a comparison between the various studies that use KGs for bias mitigation. The fairness–accuracy trade-off has been widely discussed in literature [57, 58, 59], and KG-augmented AI systems present a promising direction for addressing this challenge.

However, there are multiple researches that have shown that Knowledge Graphs are themselves prone to having biases as well [60, 61]. So, it is also crucial to make sure that KGs used for augmenting systems do not themselves contain biases. As the field matures, integrating KGs responsibly and effectively will be key to building more unbiased, equitable, and context-aware AI systems.

IX. Discussion

This study reviewed 99 publications to examine the overarching patterns of bias research and their evolution over time. Specifically, the analysis focused on research progression, institutional and cross-national collaboration, and country-level contributions across domains, considering both aggregate authorship and first-authorship patterns. The discussion is organized along five dimensions: (1) domain-specific research, its temporal evolution; (2) authorship structures; (3) domain specific trends; (4) cross-national collaboration patterns; and (5) the application of graph-based approaches to bias mitigation. Taken together, these analyses reveal the development of AI bias research over time, the strategic priorities of countries and institutions, and the disparities in leadership, collaborative engagement, and the methodological evolution driven by technological innovation.

A. Domain-Specific AI Bias Research and Temporal Evolution

Overall, we observe (see Sections VII-A and VII-B) a steady growth in AI bias and fairness research, reflecting the increasing recognition of technological challenges and the need to mitigate their potential societal impact. Much of the earlier work has concentrated on *General Fairness & Bias Mitigation* strategies, primarily because such approaches are adaptable across multiple domains. However, since 2022, there has been a notable increase in research on *LLMs & NLP*, coinciding with the release and widespread adoption of LLMs such as ChatGPT. This trend underscores both the opportunities and risks introduced by emerging innovations. In parallel, the rise of LLMs has also driven the exploration of novel methods

TABLE II
APPLIED KNOWLEDGE GRAPH–BASED BIAS MITIGATION APPROACHES

Study	**Domain / Task**	**Contribution**	**Limitations**
Contextual knowledge graph approach to bias-reduced decision support systems [20]	Decision support (license plate recognition)	Introduced contextual CKGs to guide dataset debiasing before training	Narrow evaluation domain
Detecting and mitigating bias in LLMs through knowledge graph-augmented training [21]	LLM fairness across Bias in Bios, CelebA, and COMPAS datasets	Proposed KGAT, integrating KGs via GNNs and multi-head attention, improving both fairness and accuracy	Missing results on CelebA
StereoKG: Data-Driven Knowledge Graph Construction for Cultural Knowledge and Stereotypes [55]	Cultural bias in NLP (stereotypes across religions & nationalities)	Built StereoKG of cultural stereotypes; showed intermediate MLM training on the KG improves cultural awareness and stereotype-sensitive classification	Limited group coverage (5 religions, 5 nationalities); risk of reinforcing stereotypes; ethical concerns
Visual Affect Around the World: A Large-Scale Multilingual Visual Sentiment Ontology [51]	Visual sentiment analysis (MVSO ontology capturing multilingual & cultural variations)	Used KGs for interpretability and explainability, enabling tracing of biased predictions back to knowledge base	Focused mainly on sentiment analysis; applicability to broader fairness/bias tasks remains limited
Semantics of the Black-Box: Can Knowledge Graphs Help Make Deep Learning Systems More Interpretable and Explainable? [56]	Education (student performance prediction, explainability)	Proposed a conceptual framework leveraging KGs for explainability in deep models	Mainly conceptual framework; lacks large-scale empirical validation

for bias mitigation, including *Graph-Based Fairness & Bias Mitigation*, which leverages KGs to capture context and address bias in a more systematic manner.

However, domain-specific areas such as *Health & Clinical AI* and *Recommender Systems* remain underexplored relative to their significant societal impact. Moreover, the uneven distribution of research across domains and methodologies highlights the need for greater diversification—not only in developing technical strategies but also in aligning research priorities with high-stakes application areas.

B. Domain-Specific Trends in AI Bias Research Across Institutions and Countries

The United States emerges as the dominant contributor across nearly all domains in both the all-authors and first-author (see Sections VII-C and VII-D) analyses, leading in key areas such as *General Fairness & Bias Mitigation*, *Health & Clinical AI*, *LLMs & NLP*, and *Recommender Systems*. The United Kingdom and Germany, in contrast, demonstrate notable strength in newer approaches such as *Graph-Based Fairness & Bias Mitigation*, with Germany also showing a particular concentration in *Health & Clinical AI*. Countries like Hong Kong and Italy appear to focus more narrowly on underexplored domains such as recommender systems, while China shows a stronger emphasis on newer technologies, particularly LLMs.

When the analysis is restricted to first authors, some shifts in leadership distribution become apparent. Nevertheless, the United States remains the dominant contributor, reflecting its substantial resource investment and position as a global research powerhouse. Other developed nations, such as the United Kingdom and France, also consistently occupy leading positions, underscoring both their capacity for broad contributions and their ability to channel and shape the research agenda. The United States demonstrates particular leadership in foundational domains and high-impact areas such as healthcare, highlighting its influence on research with significant societal implications. Meanwhile, the United Kingdom and Germany have emerged as leaders in evolving approaches such as graph-based methods for bias mitigation, indicating shifts in research strategies across regions. Within Asia, China and Hong Kong emerge as the primary contributors, while India and South Korea maintain only minimal presence. Overall, the region lags behind the United States and Europe in terms of research volume and leadership.

At the institutional level, the landscape remains similarly concentrated. In both all-authors and first-author perspectives, leading United States institutions—including MIT, Stanford University, Harvard University, Carnegie Mellon University, Rutgers University, the University of Pennsylvania, and the University of Washington—dominate the field. Outside the United States, only a small number of institutions such as *University College London*, *Tsinghua University*, *The Hong Kong Polytechnic University*, *Academic College of Tel Aviv-Yafo*, and the *Fraunhofer Institute for Digital Medicine* carry significant weight, often shouldering the representation of their countries. Notably, Indian institutions, despite the country's growing global profile, show a striking lack of presence in AI bias research.

C. Cross-National Collaboration Patterns in AI Bias Research

The cross-country collaboration network (see Section VII-E) reveals a concentration of partnerships among a few developed nations. The United States and the United Kingdom emerge as central hubs, forming the largest number of joint publications. The United Kingdom has a high concentration of collaboration with European countries such as Germany, Austria, and Italy and also Australia and Canada. The United States, however, shows a different pattern with high collaboration with Asia, particularly *China* and *Hong Kong*.

Overall, the collaboration patterns indicate that research on AI bias is predominantly driven by transatlantic partnerships, with Europe and North America forming the strongest ties. In contrast, Asian countries engage in comparatively fewer international collaborations, limiting their visibility and leadership in the field. This disparity reflects not only a gap in research investment across Asia but also the need for significant resource allocation and strategic initiatives to bridge it.

The uneven collaboration landscape raises critical concerns about the exclusion of perspectives from underrepresented and developing countries, which risks reinforcing biases in emerging technologies. Addressing this imbalance requires building more inclusive international collaborations, particularly through greater engagement by developed countries with institutions in the Global South, to ensure that diverse perspectives are captured and that AI systems are developed in a more equitable and representative manner.

D. Authorship Structures in AI Bias Research

When examining authorship patterns (see Section VII-F), the United States clearly dominates, contributing **228 authors across 86 papers**—by far the largest share among all countries. The United Kingdom, China, and Germany follow, though at a considerably smaller scale. Countries such as Canada, Italy, France, and Switzerland demonstrate moderate participation, while most other nations contribute only marginally. This highlights the strong concentration of research activity within a few nations. This disparity becomes even more visible under **first-authorship analysis**. The United States yet again leads, with **44 first authors across 48 papers**, reaffirming its central role not only in overall contributions but also in steering research agendas. The United Kingdom and Germany remain prominent, though less so relative to the U.S. Interestingly, China, Australia, and Canada become more competitive when considering first authorship, suggesting that while their overall contributions are smaller, they often assume leadership roles in the projects they join.

These patterns highlight both the **concentration of research leadership in developed nations** and the **underrepresentation of many regions**. Addressing this imbalance is crucial to building a more equitable global research ecosystem and ensuring the development of **fair and bias-free AI technologies**.

E. Graph-Based Approaches to Bias Mitigation

The graph based bias mitigation approaches (see Section VIII) have been applied to diverse domains. Research have demonstrated that the application of KG can mitigate biases in a way that does not reduce accuracy. We also observe from the studies [51, 56] that ontology based frameworks can enhance interpretability by tracing the biased predictions. We see that such approaches enable improvement in data processing, model training, interpretability, and representation, helping systems become aware of cultural, linguistic, and other contextual nuances that might be ignored in traditional AI pipelines.

X. Recommendations

This section outlines the recommendations emerging from our analysis and synthesis of the 99 papers reviewed.

a) Increase Bias Research in Critical Areas:: Despite notable progress and investments in AI-bias research, critical domains such as healthcare and recommender systems—which bear substantial societal implications—remain insufficiently addressed. It is therefore recommended to devote significant research attention and resources toward these domains to ensure equitable, reliable, and socially responsible AI systems.

Make Inclusive Cross-National and Institution Collaboration: Our analysis reveals significant disparities in cross-country and institutional collaborations within AI-bias research, with contributions largely concentrated among developed nations and a selected institution. This concentration risks perpetuating bias in AI systems, as they fail to adequately reflect the perspectives, contexts, and cultural realities of developing regions. To address this imbalance, it is recommended that research powerhouses—particularly the United States—actively establish equitable collaborations with developing nations, ensuring broader representation and inclusivity in shaping AI systems.

b) Do not follow the trend:: Our analysis (see Fig. 7) indicates that research activity largely follows prevailing trends—for example, LLMs and NLP—reflecting the rapid adoption and advancement of these technologies. While this focus demonstrates responsiveness to emerging developments, it also exposes an imbalance, as critical domains such as healthcare have received comparatively less attention. Research on AI bias should not be confined to a few popular areas. It is therefore recommended that, alongside emerging domains, equal priority be given to high-impact areas where biased outcomes can lead to serious and tangible societal consequences.

c) Asia needs to priotize investment in research:: Compared to the United States, Europe, and other developed regions, Asia—particularly the Global South, including India—lags significantly in AI-bias research, institutional participation, and cross-country collaborations. To address this gap, Asia must make substantial investments in both resources and strategic planning to strengthen research capacity, promote equitable technological development, and enhance its visibility in the global research landscape.

d) Use knowledge graphs for bias detection and mitigation:: Throughout this survey, we reviewed multiple experiments where KGs were used to expose biases and improve fairness - without incurring the usual trade - offs in model performance. In some cases, KG integration even improved both fairness and accuracy. However, as seen via the analytics produced by our research, Fig. 7 and Fig. 5, minimal focus is directed towards the usage of Knowledge Graphs for Bias Detection and Mitigation. Given their ability to bring context, structure, and explainability into AI systems, we strongly recommend wider adoption of KG-based mitigation techniques, especially in domains prone to representational bias, such as recommender systems, LLMs, and healthcare AI.

XI. Conclusion & Future Work

In this study, we reviewed 99 curated papers to investigate AI bias, examining research trends, cross-country collaborations, and disparities across domains and regions. We also explored the emerging role of KGs in addressing bias within AI systems. Section VII addresses research questions 1 and 2, focusing on regional representation and the evolution of AI-bias research, while Section VIII addresses research question 3 concerning the use of KGs.

One limitation of this survey is that it includes only 99 papers, representing just a fraction of the rapidly expanding AI-fairness literature. Future research should broaden the scope to include a wider range of publications and domains and a more detailed analysis of authorship and talent distribution. Future work should broaden the scope across domains, incorporate more detailed authorship analyses, and further extend the dashboard. With richer data, the dashboard could uncover how research priorities shift, where new centers of expertise are emerging, and how contributions evolve, thereby providing insights to guide policy, shape funding priorities, and foster global inclusivity in AI research.

CRediT authorship contribution statement

Abhash Shrestha: Methodology, Investigation, Software, Data Analysis, Writing – Original Draft, Validation. **Sanju Tiwari**: Investigation, Writing – Review & Editing. **Tek Raj Chhetri**: Conceptualization, Data Analysis, Validation, Methodology, Software, Investigation, Supervision, Writing – Review & Editing, Project Administration.

References

[1] Bakr Ahmed Taha, Ahmed C. Kadhim, Ali J. Addie, Adawiya J. Haider, Ahmad S. Azzahrani, Pankaj Raizada, Sarvesh Rustagi, Vishal Chaudhary, and Norhana Arsad. Advancing cancer diagnostics through multifaceted optical biosensors supported by nanomaterials and artificial intelligence: A panoramic outlook. *Microchemical Journal*, 205:111307, 2024.

[2] Amit Gangwal and Antonio Lavecchia. Unleashing the power of generative ai in drug discovery. *Drug Discovery Today*, 29(6):103992, 2024.

[3] Cyril Picard, Kristen M. Edwards, Anna C. Doris, Brandon Man, Giorgio Giannone, Md Ferdous Alam, and Faez Ahmed. From concept to manufacturing: Evaluating vision-language models for engineering design. *arXiv preprint arXiv:2311.12668*, November 2023. Preprint.

[4] Luciano A. Abriata. The nobel prize in chemistry: past, present, and future of ai in biology. *Communications Biology*, 7(1):1409, Oct 2024.

[5] Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, December 2023.

[6] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernandez, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10(3):e1356, 2020.

[7] Amit Giloni, Edita Grolman, Tanja Hagemann, Ronald Fromm, Sebastian Fischer, Yuval Elovici, and Asaf Shabtai. Benn: Bias estimation using a deep neural network. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):117–131, 2024.

[8] Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A. Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W. Bates, Raja-Elie E. Abdulnour, Atul J. Butte, and Emily Alsentzer. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22, Jan 2024.

[9] Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. Investigating bias in LLM-based bias detection: Disparities betwceen LLMs and human perception, January 2025.

[10] Adrienne Brackey and Ricardo Cortez. Analysis of racial bias in northpointe's compas algorithm, 2019.

[11] Christoph Engel, Lorenz Linhardt, and Marcel Schubert. Code is law: how compas affects the way the judiciary handles the risk of recidivism. *Artificial Intelligence and Law*, Feb 2024.

[12] Sinead O'Connor and Helen Liu. Gender bias perpetuation and mitigation in ai technologies: challenges and opportunities. *AI & SOCIETY*, 39(4):2045–2057, Aug 2024.

[13] Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 1(2):1–52, June 2024.

[14] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, July 2021.

[15] Hio Tong Pang, Xiaolin Zhou, and Mingyuan Chu. Cross-cultural differences in using nonverbal behaviors to identify indirect replies. *Journal of Nonverbal Behavior*, 48(2):323–344, Jun 2024.

[16] Shota Uono and Jari K. Hietanen. Eye contact perception in the west and east: A cross-cultural study. *PLOS ONE*, 10(2):1–15, 02 2015.

[17] Paula Reyero Lobo, Enrico Daga, Harith Alani, and Miriam Fernandez. Semantic Web technologies and bias in artificial intelligence: A systematic literature review. *Semantic Web*, 14(4):745–770, April 2023.

[18] Tek Raj Chhetri, Anelia Kurteva, Jubril Gbolahan Adigun, and Anna Fensel. Knowledge graph based hard drive failure prediction. *Sensors*, 22(3), 2022.

[19] Tek Raj Chhetri, Armin Hohenegger, Anna Fensel, Mariam Aramide Kasali, and Asiru Afeez Adekunle. Towards improving prediction accuracy and user-level explainability using deep learning and knowledge graphs:

A study on cassava disease. *Expert Systems with Applications*, 233:120955, 2023.

[20] Guang-Li Huang, , and Arkady Zaslavsky. Contextual knowledge graph approach to bias-reduced decision support systems. *Journal of Decision Systems*, 33(sup1):29–46, December 2024.

[21] Rajeev Kumar, Harishankar Kumar, and Kumari Shalini. Detecting and mitigating bias in llms through knowledge graph-augmented training. In *2025 International Conference on Artificial Intelligence and Data Engineering (AIDE)*, pages 608–613, , February 2025. IEEE.

[22] Eleni Ilkou and Maria Koutraki. Symbolic vs sub-symbolic ai methods: Friends or enemies? In *CIKM (Workshops)*, volume 2699, 2020.

[23] Varadraj Gurupur and Thomas T. H. Wan. Inherent Bias in Artificial Intelligence-Based Decision Support Systems for Healthcare. *Medicina*, 56(3):141, March 2020. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[24] Wenqi Fan, Xiangyu Zhao, Xiao Chen, Jingran Su, Jingtong Gao, Lin Wang, Qidong Liu, Yiqi Wang, Han Xu, Lei Chen, and Qing Li. A comprehensive survey on trustworthy recommender systems, 2022.

[25] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. Representation Bias in Data: A Survey on Identification and Resolution Techniques. *ACM Comput. Surv.*, 55(13s):293:1–293:39, July 2023.

[26] Zeyu Tang, Jiji Zhang, and Kun Zhang. What-is and how-to for fairness in machine learning: A survey, reflection, and perspective. *ACM Computing Surveys*, 55(13s):1–37, July 2023.

[27] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38, April 2024.

[28] Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models, 2023.

[29] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

[30] Zhibo Chu, Zichong Wang, and Wenbin Zhang. Fairness in large language models: A taxonomic survey. *ACM SIGKDD Explorations Newsletter*, 26(1):34–48, July 2024.

[31] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Qipeng Guo, Xiangkun Hu, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Xuming Hu, Zehan Qi, Wenyang Gao, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. Survey on factuality in large language models. *ACM Computing Surveys*, June 2025.

[32] Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. A survey of large language models in medicine: Progress, application, and challenge, 2023.

[33] Shiran Dudy, Thulasi Tholeti, Resmi Ramachandranpillai, Muhammad Ali, Toby Jia-Jun Li, and Ricardo Baeza-Yates. Unequal Opportunities: Examining the Bias in Geographical Recommendations by Large Language Models. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 1499–1516, Cagliari Italy, March 2025. ACM.

[34] Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 38(5):3043–3101, Sep 2024.

[35] Ashish Kumar, Vivekanand Aelgani, Rubeena Vohra, Suneet K. Gupta, Mrinalini Bhagawati, Sudip Paul, Luca Saba, Neha Suri, Narendra N. Khanna, John R. Laird, Amer M. Johri, Manudeep Kalra, Mostafa M. Fouda, Mostafa Fatemi, Subbaram Naidu, and Jasjit S. Suri. Artificial intelligence bias in medical system designs: a systematic review. *Multimedia Tools and Applications*, 83(6):18005–18057, July 2023.

[36] Shahriar Akter, Grace McCarthy, Shahriar Sajib, Katina Michael, Yogesh K. Dwivedi, John D'Ambra, and K. N. Shen. Algorithmic bias in data-driven innovation in the age of AI. *International Journal of Information Management*, 60:102387, October 2021.

[37] Pritha Bhandari. Sampling Bias and How to Avoid It | Types & Examples, May 2020.

[38] Frans J. Oort, Mechteld R. M. Visser, and Mirjam A. G. Sprangers. Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. *Journal of Clinical Epidemiology*, 62(11):1126–1137, November 2009.

[39] Kassiani Nikolopoulou. What Is Information Bias? | Definition & Examples, November 2022.

[40] Different Types of Bias in Research - CASP.

[41] Aria Khademi and Vasant Honavar. Algorithmic bias in recidivism prediction: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13839–13840, 2020.

[42] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 2145–2148, , October 2020. ACM.

[43] Yiran Du. Confirmation bias in generative ai chatbots: Mechanisms, risks, mitigation strategies, and future research directions, 2025.

[44] Mi Zhou, Vibhanshu Abhishek, Timothy Derdenger, Jaymo Kim, and Kannan Srinivasan. Bias in generative ai. arXiv preprint arXiv:2403.02726, March 2024. Preprint, not peer-reviewed.

[45] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, October 2019.

[46] Joy Buolamwini and Timnit Gebru. Gender shades:

Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, , 23–24 Feb 2018. PMLR. https://proceedings.mlr.press/v81/buolamwini18a.htmll.

[47] Joachim Baumann, Anikó Hannák, and Christoph Heitz. Enforcing group fairness in algorithmic decision making: Utility maximization under sufficiency. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2315–2326, , June 2022. ACM.

[48] Abhash Shrestha, Tek Raj Chhetri, and Sanju Tiwari. Algorithmic bias survey dashboard. https://cairnepal.github.io/algorithmic-bias-survey/, 2025. Accessed on 7/July/2025.

[49] Abhash Shrestha, Tek Raj Chhetri, and Sanju Tiwari. Source code algorithmic bias survey dashboard. https://github.com/CAIRNepal/algorithmic-bias-survey, 2025. Accessed on 7/July/2025.

[50] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations, 2020.

[51] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM international conference on Multimedia*, MM '15, pages 159–168, , October 2015. ACM.

[52] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 120–128, , January 2019. ACM.

[53] Yuanhan Zhang, Zhenfei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations, 2020.

[54] ProPublica. Propublica compas analysis. https://github.com/propublica/compas-analysis, 2016. GitHub repository.

[55] Awantee Deshpande, Dana Ruiter, Marius Mosbach, and Dietrich Klakow. StereoKG: Data-Driven Knowledge Graph Construction for Cultural Knowledge and Stereotypes, May 2022. arXiv:2205.14036 [cs].

[56] Manas Gaur, Keyur Faldu, and Amit Sheth. Semantics of the Black-Box: Can knowledge graphs help make deep learning systems more interpretable and explainable?, October 2020.

[57] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM Transactions on Software Engineering and Methodology*, 32(4):1–30, May 2023.

[58] Christina Trotter and Yixin Chen. Exploring fairness-accuracy trade-offs in binary classification: A comparative analysis using modified loss functions. In *Proceedings of the 2024 ACM Southeast Conference on ZZZ*, ACM SE '24, pages 148–156, , April 2024. ACM.

[59] Drago Plecko and Elias Bareinboim. Fairness-accuracy trade-offs: A causal perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(25):26344–26353, April 2025.

[60] Khine Myat Thwe, Teeradaj Racharak, and Minh Le Nguyen. Gender Bias Analysis in Commonsense Knowledge Graph Embeddings. In *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–6, October 2023. ISSN: 2694-4804.

[61] Styliani Bourli and Evaggelia Pitoura. Bias in Knowledge Graph Embeddings. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 6–10, December 2020. ISSN: 2473-991X.