Issues in Information Systems
Volume 25, Issue 4, pp. 491-506, 2024

DOI: https://doi.org/10.48009/4_iis_2024_138

Information system cognitive bias classifications and fairness in machine learning: Systematic review using large language models

Stephen Surles, *Dakota State University, stephen.surles@trojans.dsu.edu*
Cherie Noteboom, *Dakota State University, cherie.noteboom@dsu.edu*

Abstract

The objectives of this systematic review are to (1) gather all relevant previous works that attempt to classify known human-introduced cognitive biases and their bias reduction methods as it exists in Machine Learning (ML) in each of the three phases of the ML process – PRE-processing, the gathering of data; IN-processing, the model generation; and POST-processing, the results dissemination, and their bias reduction methods; (2) use a Large Language Model (LLM) to aid in classification of results; (3) providing a novel model for future systematic literature reviews (SLR). This work further seeks to identify the cognitive bias and methods of reduction within all phases of ML. PRISMA statement methodologies were employed to prepare this systematic review. Following these guidelines, electronic peer-reviewed sources were performed, refined, and documented producing 2107 results which were then refined to 19 works that covered the breadth of our research subject. These results showcase human-centric bias classification groupings and their mitigation methodologies identified by location within the ML process. Furthermore, the usage of a LLM proved to be an effective methodology to summarize the results of the systematic review and provided a functional methodology for performing future reviews. Two novel artifacts were introduced, (1) the ALiS framework for using LLMs to aid in the development of a Systematic Literature Review, (2) a conceptual framework for researching information systems cognitive biases and their reduction methods throughout all the phases of ML.

Keywords: bias, fairness, machine learning, large language model, cognitive

Introduction

Throughout our lives, we acquire biases shaped by experiences, interactions, and relationships, as highlighted by McLarney et al. (2021). Often, we remain oblivious to these biases, which are then reflected in our actions and the digital artifacts we create (Chouldechova & Roth, 2020). In today's digital age, where virtually every aspect of life can be captured and utilized for data-driven decisions, including automated decision-making, these biases become embedded in our systems (Mirjam et al., 2021). Consequently, the data generated can be qualitatively flawed, potentially leading to errors in both qualitative and quantitative analysis, as noted by Zhang et al., (2020). These biases not only result in fundamentally flawed data but also hinder our ability to recognize these defects. Over time, these biases get reincorporated into the datasets that inform future machine learning models, perpetuating the cycle of bias (Silva & Kenney, 2019). The paper by Harris (2020) explored methodologies to mitigate cognitive biases in Machine Learning (ML) for decision making, applying algorithmic processes to each phase of the ML process to reduce bias. However,

this paper did not explore methodologies outside of algorithms. Given the vast and complex landscape of cognitive biases and the age of this paper, we revisited this subject with a broader scope.

What is cognitive bias?

Cognitive bias is a systematic deviation in judgment and decision-making inherent to all humans, often resulting from cognitive limits, motivational aspects, or evolutionary adaptations to our environments (Wilke & Mata, 2012). Additionally, in the realm of research, bias can refer to systematic flaws in studies or analyses that result in misleading evidence (Stegenga, 2018). In the field of Information Systems, cognitive biases are classified into eight categories: (1) Perception Biases, (2) Pattern Recognition Biases, (3) Memory Biases, (4) Decision Biases, (5) Action-oriented Biases, (6) Stability Biases, (7) Social Biases, and (8) Interest Biases. Each category encompasses one or more specific biases that may influence Machine Learning applications, as shown in Table 1 (Fleischmann et al., 2014).

Table 1: Cognitive Bias Categories in IS (Fleischmann et al., 2014)

Category	Biases
Perception biases	framing, negativity bias, halo effect, selection bias, representativeness bias, sequential bias, priming effect, recency effect, biased perception of partitioned prices, emotional bias, primacy effect, selective perception
Pattern Recognition biases	confirmation bias, availability bias, reasoning by analogy, disconfirmation bias
Memory biases	reference point dependency
Decision biases	irrational escalation, reactance, illusion of control, cognitive dissonance, mental accounting, mere exposure effect, exponential forecast bias, ambiguity effect, zero-risk bias, input bias, base-rate fallacy, omission bias
Action-orientated biases	overconfidence, optimism bias
Stability biases	anchoring, sunk cost bias, status-quo bias, loss aversion, endowment effect
Social biases	herding, stereotype, value bias, attribution error, cultural bias
Interest biases	after-purchase rationalization, self-justification

What is fairness in machine learning?

Fairness in machine learning is conceptualized primarily through two frameworks: statistical and individual fairness (Chouldechova & Roth, 2020). Traditional research often emphasizes statistical fairness, which involves equalizing outcomes across both protected and non-protected groups to achieve demographic parity (Dwork et al., 2011). On the other hand, individual fairness focuses on treating similar individuals in a comparable manner and ensuring that dissimilar individuals are not treated alike – hoping to reach the goal of consistent outcomes for those who are similar (Friedler et al., 2021; García-Soriano & Bonchi, 2020).

What are the phases of machine learning?

The lifecycle of machine learning is segmented into three phases: PRE, IN, and POST. The PRE phase centers on data preparation, including the selection of training data, where various forms of bias such as institutional, individual, and sampling biases can be introduced (Bacelar, 2021). The IN phase typically introduces algorithmic biases that encompass feature selection, algorithm development, and model selection, during which assumptions are made (Yapo & Weiss, 2018). Lastly, the POST phase involves validating the model, synthesizing results, and disseminating findings.

Exploring the Gap

Despite the growth in cognitive bias research over the past two decades, there remains a significant gap in studies applying bias reduction methods to Machine Learning (Fleischmann et al., 2014; Kliegr et al., 2021). This study aims to bridge this gap by critically reviewing existing literature on biases in machine learning. In exploring this gap, we propose the following research questions.

Research Question 1: Are cognitive biases present in all the machine learning phases outside of just algorithmic processes, and if so, what are the methods of mitigation that exist to increase fairness?

The deployment of machine learning (ML) systems across various domains, including healthcare, finance, and criminal justice, has highlighted the critical issue of cognitive biases in these systems. Biases in ML can lead to unfair, inaccurate, and potentially harmful outcomes, reinforcing existing social inequities and introducing new forms of discrimination. Despite substantial advancements in bias mitigation techniques, there is often a gap in the research between identifying a specific cognitive bias and directly correlating it with the most effective remediation method. This gap can hinder the development and implementation of fairer and more reliable ML systems.

Understanding how cognitive biases are mitigated in ML, even in the absence of direct correlations between biases and remediation methods, is critical for several reasons – there needs to be broad applicability, improved fairness and accuracy, guidance on future research, and guidance on policy and regulation.

Consequently, a conceptual framework was developed that maps Cognitive Biases (Cognitive Phase) to their Processing (Application Phase) and Debiasing (Methodology Phase) strategies. This framework examines how these strategies impact the Cognitive Phase (Figure 1).

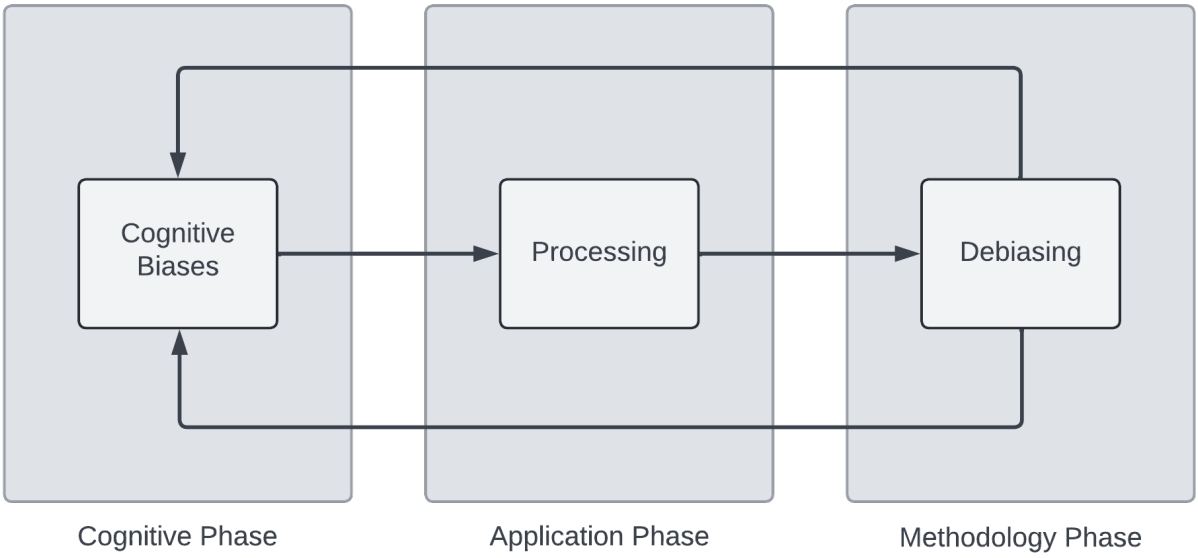

Figure 1: Phase-Based Cognitive Bias Classification Framework

Research Question 2: How can the usage of Large Language Models aid the researcher in the application of a conceptual framework during the systematic literature review process in the area of cognitive bias?

Why involve a Large Language Model (LLM) in analyzing these papers and shaping the findings? The demand for systematic reviews and meta-analyses is increasing, as is the quantity of available literature

(Ebrahim & Huffman, 2022). Utilizing LLMs enables researchers to synthesize a much larger volume of information more efficiently than ever before.

The systematic literature review (SLR) process is fundamental for synthesizing existing research and identifying gaps in knowledge across various fields. In the study of cognitive biases, a comprehensive SLR can be particularly challenging due to the interdisciplinary nature of the topic, the vast amount of literature, and the nuanced definitions and applications of different biases. LLMs offer a promising tool for researchers to enhance the efficiency and effectiveness of the SLR process by assisting in the application of conceptual frameworks. By providing efficiency in data handling, enhanced accuracy and consistency, integration of interdisciplinary knowledge, and identification of research gaps, LLMs have the potential to add rigor and velocity to SLRs simultaneously.

Recent advancements in LLMs have shown their potential to transform this process by providing tools for automated literature searches, data extraction, and synthesis. However, the specific application of LLMs in the context of cognitive bias research, particularly in applying conceptual frameworks, remains underexplored.

Methodology

The starting point of this review followed the PRISMA statement methodology for performing systematic reviews (Liberati et al., 2009). An electronic search of ACM Digital, IEEE Xplore, ProQuest, Science.gov, and Web of Science was conducted to identify relevant literature through April 2024. No results were discarded due to age of publication if relevant to the subject. All results were required to be from peer reviewed sources and not in pre-print status unless accepted for publication. Additional searches were performed based on reference lists in papers identified to have relevant material.

In developing the search strategy, keywords and operators were tested to produce the most inclusive results while reducing the highest quantity of false positives. Bias can have many different definitions and usages, even in the realm of machine learning. Because the subject of this work is specific to the biases that humans introduce in the three phases of machine learning statistical biases such as sampling and estimation were excluded.

The search strategy employed used fairness and bias interchangeably, if possible, while also excluding statistical keyword(s). When possible, searches were only performed on title and abstract. Full text searches were excluded to focus only on work that explicitly includes the relevant concepts and to limit the results. All searches used the terms "bias" and "fairness" interchangeably and mandated their inclusion. Additionally, the term "machine learning" was necessary. Any results containing "statis*" with a wildcard were excluded to omit papers focused on statistics, as they are outside the scope of this review. The following search term was used, with site specific features to reduce results to peer reviewed results, in English.

(Bias OR Fairness) AND “Machine Learning” AND NOT Statis*

When selecting papers for this review, it was critical we applied a selection criterion to our results that sought to address our research question directly, following the phase-based cognitive bias classification conceptual framework outlined above. This framework guided the formulation of the following questions when evaluating our results for inclusion:

1. Did the study specifically identify bias or discrimination within the context of machine learning?
2. Did the research address or imply the influence of human cognitive biases?
3. Was there an effort to mitigate biases during the machine learning phases?
4. Did the bias mitigation effort consider fairness as a desired outcome?

Only papers that discussed bias or fairness in machine learning were included when they also explicitly mentioned or implied some form of cognitive bias as required by the research guidelines set forth. Duplicates were excluded, as well as papers where full text was not available. Papers that did not have a methodology for removing or mitigating bias were also rejected for final inclusion.

It is important to employ our conceptual model in crafting prompts for the LLM. By providing clear and concise instructions, the LLM is best prepared to provide the most accurate results possible. The AI-enhanced Literature Systematic Review Framework (ALiS) was developed to use LLMs in conjunction with existing systematic literature review methodologies (Figure 2).

To facilitate prompt engineering, prompts were crafted using the following steps, aligned to three distinct phases, Model Training, Model Application, and Analysis. ChatGPT-4o was the Large Language Model used based on overall accuracy of the currently available LLMs (Polak & Morgan, 2024), and our comparative analysis of output between OpenAI models.

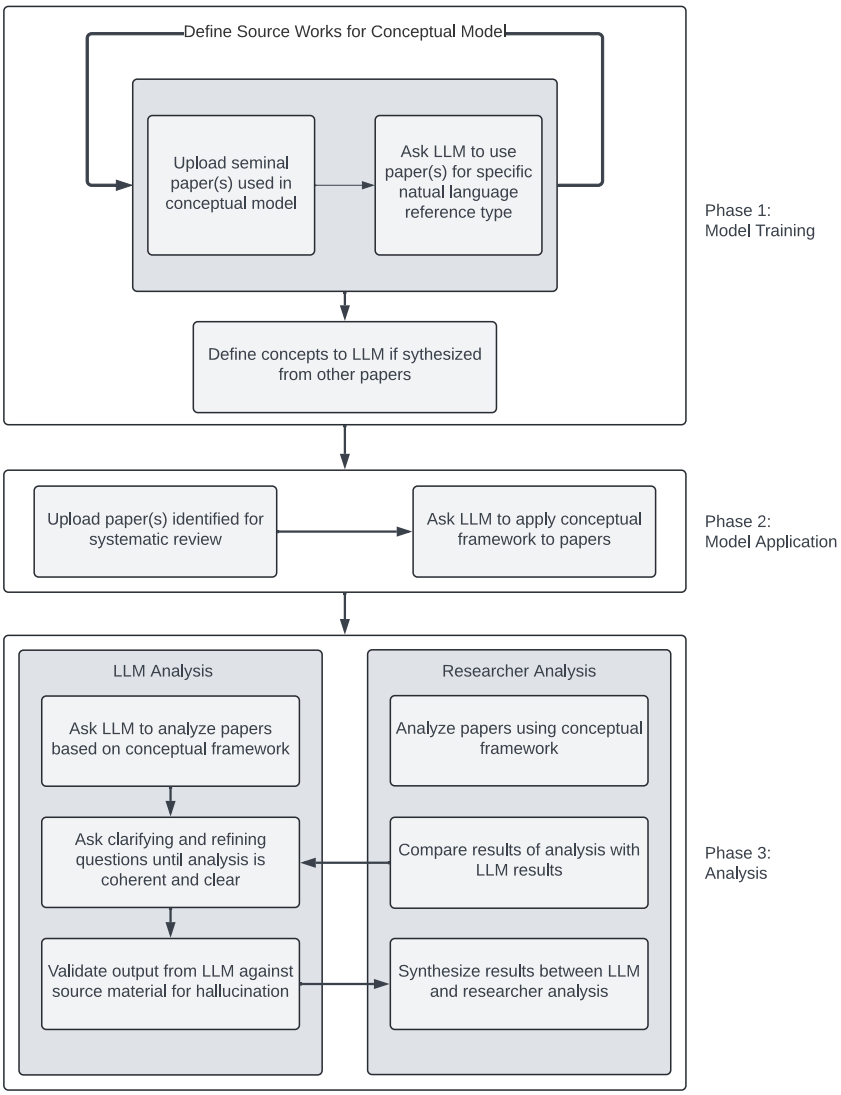

Figure 2: AI-enhanced Literature Systematic Review Framework (ALiS)

When utilizing the ALiS framework the following steps should be employed to achieve the highest quality outcome. During the model training phase, the researcher prepares the LLM by creating artifacts in memory that the LLM will reference later for retrieval-augmented generation. This step provides a strong foundation for the results and allows the definition of core concepts necessary to perform the rest of the analysis (Lewis et al., 2020).

At the model application phase, papers are uploaded in batches. We found that including more than five papers in each session degraded the quality of the results. It is here that the LLM is asked to apply the concepts created in memory from the training phase to the uploaded papers, thereby directly associating the papers with the core concepts, and preparing the model to produce the analysis output. Below is a sample prompt used during this stage.

> *"When providing answers, please limit cognitive biases to the ones found in the Fleischmann paper and please apply the following conceptual framework: cognitive bias, processing phase (pre, in, post), and debiasing methodology. Please state in paragraph form."*

Finally, with the analysis phase, we recommend a parallel approach to analyzing the papers – one that utilizes the LLM and one that utilizes the researcher. Both paths take the approach of analyzing the papers to the conceptual framework. While the LLM utilizes prompt engineering for this purpose, the researcher performs analysis by reading. Here we must be cautious to continuously ask and validate the LLM iteratively until there is agreement by the researcher that the results from the LLM are accurate and validated. We find that this process gives additional inflection points for consideration and analysis and has the opportunity to produce higher quality output. Provided is a sample prompt for this stage, showing how the question directly follows the framework description.

> *"What are the cognitive bias categories that are addressed in each of these papers, which phase are they in, and what are the debiasing techniques?"*

Model Training – Step 1:

1. upload conceptual model paper(s)
2. define conceptual model concepts to LLM

Model Application – Step 2:

1. upload papers identified for systematic review – max 5
2. prompt LLM to apply conceptual framework elements to papers identified for systematic review

Analysis – Step 3:

1. LLM: prompt to analyze papers based on conceptual framework one at a time
2. LLM: prompt with clarifying questions until analysis is cohesive and clear
3. LLM: validate output from LLM against source material for hallucination
4. Researcher: analyze papers using conceptual model framework
5. Researcher: compare results of analysis with LLM results
6. Researcher: synthesize results between LLM and researcher analysis

Results

A total of 2107 results were identified using the search strategy described above. Promising results, based on title and abstract, were retained for full text inclusion. After de-duplication, title, and abstract filtering, 117 results remained for assessment against the eligibility criteria. There were two reports unavailable at this stage. In the full-text stage, a further 88 papers were excluded due to invalid requirements matching

against our research questions above, as well as 8 papers that were inaccessible producing the result of 19 papers as the subject of this review (see Figure 3 for PRISMA flow diagram).

The following results are organized by paper, according to our conceptual framework, listing the potential cognitive biases mitigated, the phase the mitigation occurs, and the novel mechanism in which the paper introduced fairness into the model. Results are alphabetical by author name.

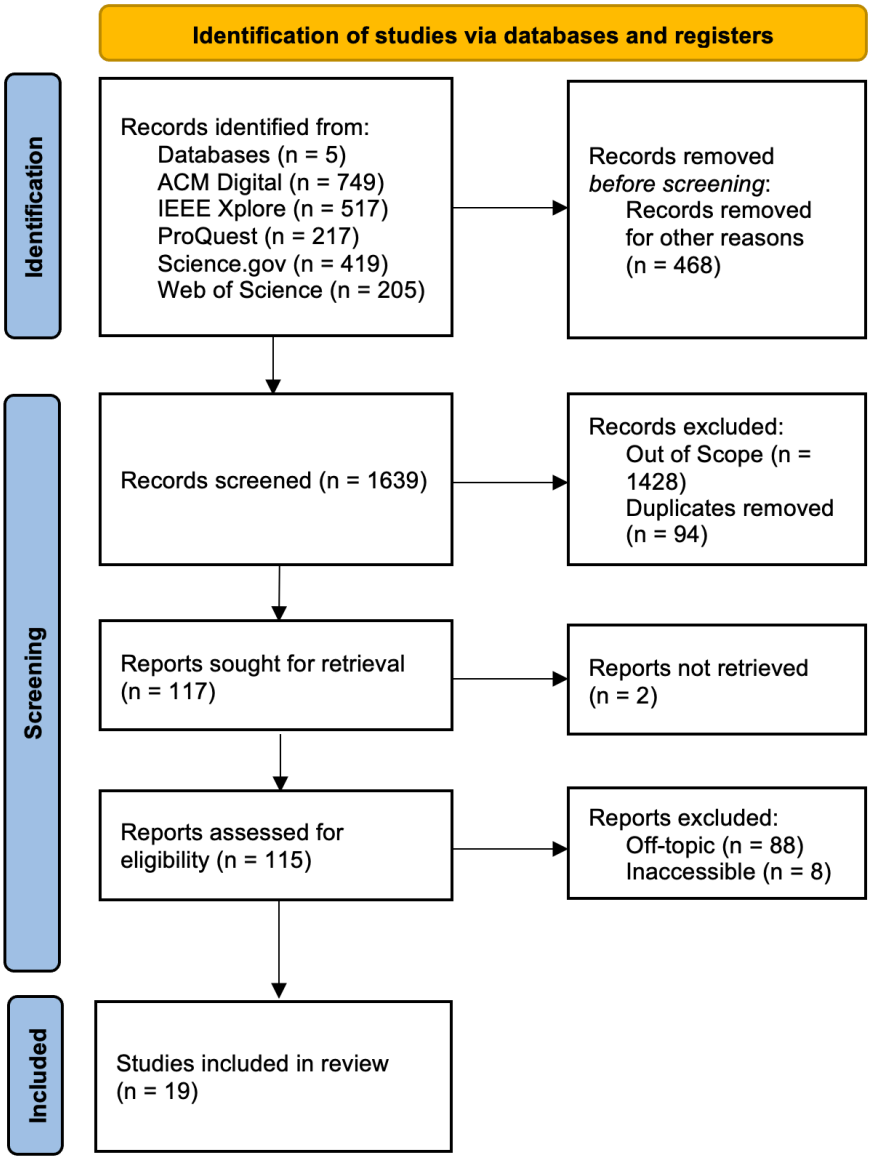

Figure 3: PRISMA Statement

The systematic review of the selected papers reveals insightful intersections among cognitive biases within the distinct phases of ML. The papers collectively aim to address and mitigate various forms of bias that may impact algorithmic decision-making processes and outcomes. The selected papers address various cognitive biases through distinct processing phases and debiasing methodologies as outlined in the Fleischmann et al. (2014) framework. These results are generated by the output of the LLM with inputs and edits by the authors.

The paper "ABCinML: Anticipatory Bias Correction in Machine Learning Applications", addresses stability biases due to changing input distributions over time (Almuzaini et al., 2022). They focus on multiple processing phases, including PRE-processing, IN-processing, and POST-processing. Their debiasing methodologies involve dynamic learning, which continuously retrains models with new data batches to address emerging biases, robust learning to develop models resilient to future changes, and anticipatory dynamic learning, which uses predictions about future data distributions to preemptively adjust model parameters and mitigate bias before it occurs.

Ashokan and Haas (2021), in their study "Fairness Metrics and Bias Mitigation Strategies for Rating Predictions", explore perception biases, pattern recognition biases, and social biases. These include selective perception from data recording methods, reference point dependency favoring more popular items or users, and cultural bias against specific groups such as gender or race. They address these biases in the PRE-processing phase and IN-processing phase. Their debiasing techniques include pre-processing to correct biases in the dataset before training, IN-processing to incorporate fairness constraints directly into learning algorithms, and POST-processing to adjust final model predictions to ensure fairness without altering the underlying model.

In the context of "Missing the missing values, The ugly duckling of fairness in machine learning", Fernando et al. (2021) find that missing values are often linked to protected attributes, advocating for imputation over discarding data to maintain fairness within the PRE-processing phase. Applying a framework to address these biases systematically improves fairness in machine learning, ensuring targeted debiasing at each processing phase.

In "On The Impact of Machine Learning Randomness on Group Fairness," the authors investigate stability biases, particularly the high variance in fairness measures due to data reshuffling during training (Ganesh et al., 2023). Their focus is on the IN-processing phase. They propose debiasing methodologies such as controlling the data order during training and implementing single-epoch adjustments. These techniques aim to minimize the impact of randomness on group fairness and improve model fairness efficiently with minimal impact on overall accuracy.

García-Soriano and Bonchi's (2020) paper on fair-by-design matching targets action-oriented and stability biases during the IN-processing phase. Their approach involves designing algorithms that inherently prevent biased outcomes by incorporating fairness distribution, ensuring that the algorithms remain fair over time and across different scenarios.

The paper "D-BIAS: A Causality-Based Human-in-the-Loop System," addresses decision biases and social biases (Ghai & Mueller, 2023). The authors examine algorithmic bias introduced by the algorithm itself and overconfidence from causal relationships in the data reflecting societal inequalities. They focus on the IN-processing phase and POST-processing phase. Their debiasing methodologies include a human-in-the-loop approach, engaging human experts to identify and mitigate biases using domain knowledge, causal modeling to understand and manipulate causal relationships that contribute to bias, and interactive visualization tools for users to adjust the causal model to reduce bias.

Hauptmann et al.'s (2023) work on maximal representative subsampling (MRS) focuses on mitigating selection and social biases during the PRE-processing phase. MRS iteratively removes or adapts instances from biased datasets to align them with representative datasets, ensuring that the training data is representative of the target population. This approach addresses pattern recognition biases related to unrepresentative sampling and reduces social biases by ensuring fair representation of underrepresented groups.

In "Parity-based Cumulative Fairness-aware Boosting" by Iosifidis et al. (2022), pattern recognition and decision biases are tackled during the IN-processing phase. The AdaFair algorithm dynamically adjusts instance weights during each boosting round to ensure fairness and balanced error rates, thus mitigating biases by considering both predictive performance and fairness at each stage.

Johndrow and Lum's (2019) paper on removing sensitive information addresses perception and decision biases in the PRE-processing phase. Their method removes sensitive information such as race from the dataset before training, ensuring that the model cannot use these attributes to make biased predictions. This PRE-processing step effectively prevents biased decision-making based on protected attributes.

Kamiran and Calders' (2012) work on data preprocessing techniques for classification without discrimination addresses decision and pattern recognition biases in the PRE-processing and IN-processing phases. They propose methods such as suppression of sensitive attributes, dataset massaging, and re-weighting or resampling to create fair training data. These techniques ensure that the training data does not perpetuate existing biases, thereby mitigating decision and pattern recognition biases.

The work "Designing Ethical Algorithms" addresses interest biases in the PRE-processing phase, specifically in training data biases (Martin, 2019). The paper discusses the ethical concerns of algorithms biased by the interests of developers or organizations. To mitigate these biases, the implementation of ethical frameworks and guidelines during development is recommended, ensuring diverse perspectives are considered. Additionally, transparency and accountability mechanisms are recommended to reduce interest biases, ensuring a more ethical approach to algorithm design.

In their paper "Toward Involving End-users in Interactive Human-in-the-Loop AI-Fairness" the authors tackle decision biases in the IN-processing phase (Nakao et al., 2022). By involving end-users in the machine learning process, the authors aim to mitigate decision biases through the incorporation of human judgment and contextual knowledge. Human-in-the-loop approaches allow end-users to interact with and influence the machine learning process, reducing biases from automated decision-making. This method ensures that human insights and contextual understanding enhance the fairness and accuracy of algorithmic outcomes.

In the work "Data Augmentation for Fairness-Aware Machine Learning" by Pastaltzidis et al. (2022), stability biases in the IN-processing phase are tackled by ensuring fairness and robustness in machine learning models trained on biased datasets. The authors use data augmentation techniques to balance the datasets, thereby reducing the impact of inherent biases. By artificially expanding the training data, the models learn from a more diverse set of examples, promoting stability and reducing bias in model performance.

In the paper "Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning " social biases, categorized as interpretation biases in the POST-processing phase, are addressed by exploring how different mathematical definitions of fairness align with human perceptions (Srivastava et al., 2019). The study finds that demographic parity closely matches people's ideas of fairness in contexts like criminal risk assessment and medical prediction. To mitigate these biases, the researchers use adaptive experiments that minimize cognitive load, helping participants make informed decisions by aligning mathematical fairness definitions with human perceptions.

Taniguchi et al. (2018) address pattern recognition biases in the IN-processing phase in their paper "A Machine Learning Model with Human Cognitive Biases Capable of Learning from Small and Biased Datasets." They leverage human cognitive biases to enhance machine learning models' ability to generalize

from small datasets. The introduction of Loosely Symmetric Naïve Bayes (LSNB) and Enhanced LSNB (eLSNB) models helps incorporate these biases, improving learning efficiency and adjusting feature weights to mitigate potential errors, thus enhancing performance with small and biased samples.

Wang et al. (2022) tackles social biases in the PRE-processing phase using an intersectional approach. This involves incorporating fairness-aware algorithms and extensive testing across different demographic groups to identify and mitigate biases early in the data preparation and model training stages. By considering multiple dimensions of identity such as race and gender together, and thoughtful domain knowledge, this methodology aims to address social biases comprehensively.

The paper by Wang et al. (2023) addresses social biases in the PRE-processing phase by incorporating multisource data, including demographic, clinical, genetic factors, and cognitive scores. This approach ensures that models are trained on a balanced and representative dataset, reducing biases across subgroups such as gender, age, and race, leading to more accurate and fair predictions.

Zhang et al. (2020) focuses on decision biases in the PRE-processing phase. Their proposed framework involves pseudo labeling to predict labels for unlabeled data, re-sampling to create fair datasets, and ensemble learning. These techniques collectively ensure equal representation of all groups and improve the accuracy and fairness of the models from the start, effectively mitigating decision biases.

Lastly, Zhang et al. (2023) deals with action-oriented biases in the IN-processing phase through the iFlipper technique. This technique involves flipping labels to improve individual fairness, ensuring that the model's decisions are fair at the individual level. By adjusting predictions in real-time, the iFlipper technique maintains fairness across different groups, addressing action-oriented biases effectively.

The reviewed papers demonstrate various approaches to addressing cognitive biases in machine learning, categorized by the processing phases (PRE-processing, IN-processing, and POST-processing) and debiasing methodologies as outlined in the Fleischmann et al. framework.

These papers collectively showcase diverse strategies for mitigating cognitive biases across different phases of the machine learning process, promoting fair and equitable outcomes. By leveraging various debiasing methodologies, these studies contribute to the development of more transparent, accountable, and ethical machine learning systems.

Conclusion

Through these collective efforts, each paper contributes to a comprehensive understanding of cognitive biases and their mitigation in machine learning. By addressing these biases at every stage, from data collection and preprocessing to algorithmic processing and deployment, these studies pave the way for developing machine learning systems that are not only technically robust but also socially equitable. This holistic approach ensures that machine learning models serve all segments of society fairly, upholding the principles of justice and integrity in their design and application. By continuously evolving and refining these methodologies, the field of machine learning can move towards a future where technology empowers rather than discriminates, fostering a more inclusive and fair society.

This narrative underscores the multifaceted nature of cognitive biases and the nuanced strategies required to address them. Each study provides a piece of the puzzle, contributing to a broader effort to build machine learning systems that are as unbiased and fair as possible. Through ongoing research and practical application of these findings, we can work towards machine learning technologies that truly benefit all of humanity.

PRE-processing techniques and data augmentation are crucial for addressing biases inherent in the training data. By modifying and balancing the data before it is used for training, these methods ensure that underrepresented groups are adequately represented, minimizing biases such as mental accounting, mere exposure effect, and selection bias.

Fairness-aware algorithms and boosting techniques embed fairness constraints within the algorithms themselves, ensuring equitable treatment of different groups. These approaches help reduce biases during the decision-making process, promoting fair and balanced outcomes.

Dynamic learning strategies and human-in-the-loop systems provide continuous updates and real-time feedback to the model, allowing for iterative bias correction. By incorporating human insights and continuously adjusting the model based on new data, these methods ensure that the model remains adaptable and responsive to changing conditions, addressing biases such as self-justification and sequential bias.

Ethical algorithm design involves incorporating ethical considerations into the development process from the outset. By involving ethicists and domain experts, potential biases can be anticipated and addressed proactively, ensuring that the model's development is guided by ethical principles and promotes transparency and accountability.

Overall, these debiasing methods provide a comprehensive framework for identifying and mitigating cognitive biases in machine learning systems, ensuring fair and unbiased outcomes. By targeting different stages of the machine learning lifecycle and addressing specific biases, these methods help promote a balanced and ethical approach to machine learning model development.

Discussion

Cognitive Bias

In answering our first research question, “how are cognitive biases in machine learning mitigated, even when the research does not directly correlate the association between the bias itself and the remediation method?”, we discovered that, while associations do exist between cognitive bias and mitigation techniques the current research is lacking in depth. The small quantity of results shows that there is a dearth of information on this subject, given how prevalent the conversation around bias in ML is. Although there is much research published about bias and fairness in machine learning, the results of this systematic review clearly show that the inclusion of cognitive bias is not often considered in those discussions. Additionally, there are not many authors who have identified ways to mitigate those biases for fairness. Although 88 papers were rejected because they did not have a specific discernable debiasing method mentioned, they were indicative of an overall pattern – one of a lack of methods. Many of these papers discussed the importance of researcher manual intervention in the ML process centered around fairness and bias concepts yet provided no guidance.

It’s clear from the emerging research that this is a very notable topic, however very few results are available that are peer reviewed or not in conference proceedings. In A Survey on Bias and Fairness in Machine Learning, Mehrabi et al. (2019), found twelve works, with only one available from a peer reviewed journal. Out of the total works reduced during full-text analysis, most recognized the need for additional research in this area. It’s encouraging that the trend for papers is shifting upwards, with the last two years having more papers than all the previous years combined.

By categorizing biases and interventions across the ML phases, this review highlights the intricate ways in which biases can be integrated into and mitigated within the phases of machine learning development and

application. The collective findings suggest that proactive measures in the design and implementation phases of algorithms are crucial for minimizing the impact of cognitive biases on algorithmic decisions, thereby fostering systems that are both ethically responsible and technologically robust.

Our findings compellingly demonstrate that beyond the algorithmic strategies for bias reduction outlined by Harris (2020), there exists a multitude of additional opportunities to enhance fairness throughout the machine learning ecosystem. These opportunities arise from a variety of sources, including but not limited to, data collection methods, model training techniques, and the deployment practices of machine learning systems.

By expanding our focus beyond algorithms alone, we can explore a broader array of interventions that might contribute to the creation of more equitable and fair machine learning applications. This broader perspective enables us to identify and leverage various points of intervention where fairness can be significantly impacted, potentially leading to more robust and universally fair ML solutions.

Large Language Models

In answering our first research question, "how can the usage of Large Language Models aid the researcher in the application of a conceptual framework during the systematic literature review process in the area of cognitive bias?", we discovered that usage of an LLM was a double-edged sword – simultaneously helpful in assessing concepts and themes within multiple works, while generating results that weren't precisely what was requested via the prompt.

The LLM results were prone to hallucination and misrepresentation, especially if more than 5 papers were included in any given analysis scenario. We found that by asking questions around the conceptual framework to papers individually provided the best and most accurate results – grouping papers and asking more narrative questions produced less satisfactory results.

Incorporating the model for using LLMs provided a valuable methodology for synthesizing the papers. However, the LLM did not provide always provide cohesive output through many iterations of prompt engineering. It was necessary for the authors to perform an analysis of the output and re-organize it in a way that followed the conceptual model of the systematic review, even though the prompts incorporated that model within its responses.

When evaluating the output of the LLM results using the ALiS framework, the authors tracked changes through the LLM produced output. Those changes were classified as either grammatical errors or understanding errors. Grammatical errors were ones that produced output that needed rewording but did not change the meaning of the output. Understanding errors fundamentally changed the meaning of the output. For analysis purposes of changes, words that were contiguous were counted as a single edit. Formatting preferences, such as changing of "pre-processing" to "PRE-processing" were not considered as a failure of the model output and are not included in the calculations.

As shown in Table 2, of 1449 words produced by the LLM output, there were 5 grammatical errors and 12 understanding errors, a success rate of 99.45% and 98.69% respectively, and a total error rate of 1.86%. Interestingly enough, the most common error produced by the LLM was omission of part of the title of the source paper.

Also of note, was the LLM was better at paraphrasing and summarizing high level concepts than it was at providing technical understanding of the articles used for analysis.

Table 2: ALiS Error Rate

Generated Word Count	Error Type	Error Count	Error Rate	Success Rate
1449	Understanding	19	0.0131	0.9869
	Grammatical	8	0.0055	0.9945

Given this outcome, it's apparent that the usage of LLMs is helpful, but does not replace the insight and understanding of the author, whom should be familiar with the subject being discussed and capable of providing discernment and judgement on applicability of LLMs. While performing systematic literature reviews with a single author is known to be rigorous using the PRISMA statement, adding in the LLM as a "second" researcher greatly enhanced the quality of the output. Often the LLM would notice themes or insights that challenged the researchers to re-evaluate their initial impressions, significantly enhancing the final product.

Limitations

There are some limitations to this work, starting with the tremendous number of cognitive biases recognized today; at the time of this paper over 100 have been identified (Blawatt, 2016). It would be very difficult to correlate all those biases directly to machine learning without the usage of a platform like an LLM. This review does not attempt to delineate all the issues with bias in ML as it exists in the many different techniques at each phase, or the specific types of ML methodologies used. These deeper classifications are potential subjects for future research based on the higher-level findings of this work.

The ALiS framework was developed and refined using the sources of this paper, on English language texts only. It is unknown of the framework is generalizable to other research domains. Based on our experience using ALiS, we believe that the framework is extendable to performing SLRs where the subject matter is, or research questions are confined within the natural language paradigm.

Lastly, there is not currently a published methodology for using LLMs to perform systematic reviews. While this paper presents an operational framework for performing this task, more research is needed in this area. However, by utilizing the power of a LLM and comparing the results with the researcher in an iterative method, validity and reliability of results are increased, providing increased rigor.

Contributions and Future Research Opportunities

The practical contributions of this paper are first, the identification of cognitive biases most found within machine learning. By identifying these biases, ML model builders can have a focus area likely to have the most impact on reduction. Second, those biases are then categorized where they may be found within the ML phases, and third, what methods exist, to reduces those biases and introduce fairness.

The research contributions of this paper are the introduction of a novel model for using LLMs to perform systematic reviews in future research, as well as some of the difficulties and guidelines necessary to make that endeavor effective. Usage of an LLM using the ALiS framework has the potential to increase rigor and validity within results as long as practitioners exercise caution in validating the results of the output. Opportunities exist for future research into an expansion of methods to reduce human cognitive bias in machine learning. In the application of LLMs in systematic reviews, more research is needed to better define prompt engineering methodologies to increase the usefulness of the LLM in performing more of the review for practitioners with a reduced or limited understanding of the source materials and concepts.

References

Almuzaini, A. A., Bhatt, C. A., Pennock, D. M., & Singh, V. K. (2022). ABCinML: Anticipatory Bias Correction in Machine Learning Applications. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1552–1560). Association for Computing Machinery.

Ashokan, A., & Haas, C. (2021). Fairness metrics and bias mitigation strategies for rating predictions. *Information Processing & Management*, *58*(5), 1. ABI/INFORM Collection.

Bacelar, M. (2021). *Monitoring bias and fairness in machine learning models: A review* [Preprint].

Blawatt, K. R. (2016). Appendix A: List of Cognitive Biases. In *Marconomics* (pp. 325–336). Emerald Group Publishing Limited.

Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, *63*(5), 82–89.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness Through Awareness. *arXiv:1104.3913 [Cs]*.

Ebrahim, S., & Huffman, M. D. (2022). Future for Systematic Reviews and Meta-Analysis. In M. Egger, J. P. T. Higgins, & G. Davey Smith (Eds.), *Systematic Reviews in Health Research* (1st ed., pp. 463–479). Wiley.

Fernando, M., Cèsar, F., David, N., & José, H. (2021). Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems*, *36*(7), 3217–3258.

Fleischmann, M., Amirpur, M., Benlian, A., & Hess, T. (2014). COGNITIVE BIASES IN INFORMATION SYSTEMS RESEARCH: A SCIENTOMETRIC ANALYSIS. *Tel Aviv*, 23.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (Im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, *64*(4), 136–143.

Ganesh, P., Chang, H., Strobel, M., & Shokri, R. (2023). On The Impact of Machine Learning Randomness on Group Fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1789–1800). Association for Computing Machinery.

García-Soriano, D., & Bonchi, F. (2020). Fair-by-design matching. *Data Mining and Knowledge Discovery*, *34*(5), 1291–1335.

Ghai & Mueller. (2023). D-BIAS: A Causality-Based Human-in-the-Loop System for Tackling Algorithmic Bias. *IEEE Transactions on Visualization and Computer Graphics*, *29*(1), 473–482.

Harris, C. G. (2020). Mitigating Cognitive Biases in Machine Learning Algorithms for Decision Making. *Companion Proceedings of the Web Conference 2020*, 775–781.

Hauptmann, T., Fellenz, S., Nathan, L., Tüscher, O., & Kramer, S. (2023). Discriminative machine learning for maximal representative subsampling. *Scientific Reports*, *13*(1), 20925.

Iosifidis, V., Roy, A., & Ntoutsi, E. (2022). Parity-based cumulative fairness-aware boosting. *Knowledge and Information Systems*, *64*(10), 2737–2770. ABI/INFORM Collection.

Johndrow, J. E., & Lum, K. (2019). An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *The Annals of Applied Statistics*, *13*(1).

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33.

Kliegr, T., Bahník, Š., & Fürnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, *295*, 103458.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *Annals of Internal Medicine*, 30.

Martin, K. (2019). Designing Ethical Algorithms. *MIS Quarterly Executive*, 129–142.

McLarney, E., Gawdiak, Y., & Nikunj, O. (2021). *NASA Framework for the Ethical Use of Artificial Intelligence (AI)* (Technical Memorandum TM-20210012886; p. 35). NASA Langley Research Center.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *arXiv:1908.09635 [Cs]*.

Mirjam, P., Nathalie, K., Prainsack, B., & Link to external site, this link will open in a new window. (2021). Not all biases are bad: Equitable and inequitable biases in machine learning and radiology. *Insights into Imaging*, *12*(1).

Nakao, Y., Stumpf, S., Ahmed, S., Naseer, A., & Strappelli, L. (2022). Toward Involving End-users in Interactive Human-in-the-loop AI Fairness. In *ACM Trans. Interact. Intell. Syst.* (Vol. 12, Issue 3, p. Article 18). Association for Computing Machinery.

Pastaltzidis, I., Dimitriou, N., Quezada-Tavarez, K., Aidinlis, S., Marquenie, T., Gurzawska, A., & Tzovaras, D. (2022). Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2302–2314). Association for Computing Machinery.

Polak, M. P., & Morgan, D. (2024). Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, *15*(1), 1569.

Silva, S., & Kenney, M. (2019). Algorithms, platforms, and ethnic bias. *Communications of the ACM*, *62*(11), 37–39.

Srivastava, M., Heidari, H., & Krause, A. (2019). Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2459–2468). Association for Computing Machinery.

Stegenga, J. (2018). *Care and Cure*. University of Chicago Press.

Taniguchi, H., Sato, H., & Shirakawa, T. (2018). A machine learning model with human cognitive biases capable of learning from small and biased datasets. *Scientific Reports (Nature Publisher Group)*, *8*, 1–13.

Wang, A., Ramaswamy, V. V., & Russakovsky, O. (2022). Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 336–349). Association for Computing Machinery.

Wang, R., Chaudhari, P., & Davatzikos, C. (2023). Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(6), 1. Research Library.

Wilke, A., & Mata, R. (2012). Cognitive Bias. In *Encyclopedia of Human Behavior* (pp. 531–535). Elsevier.

Yapo, A., & Weiss, J. (2018). *Ethical Implications of Bias in Machine Learning*. Hawaii International Conference on System Sciences.

Zhang, H., Tae, K. H., Park, J., Chu, X., & Whang, S. E. (2023). iFlipper: Label Flipping for Individual Fairness. In *Proc. ACM Manag. Data* (Vol. 1, Issue 1, p. Article 8). Association for Computing Machinery.

Zhang, T., Zhu, T, Li, J., Han, M., Zhou, W., & Yu, P. (2020). Fairness in Semi-supervised Learning: Unlabeled Data Help to Reduce Discrimination. *IEEE Transactions on Knowledge and Data Engineering*, 1–1.