Human Cognitive Bias Mitigation Approaches to Fairness within the Machine Learning Value Chain: A Review and Research Agenda

Stephen Surles
Dakota State University
stephen.surles@trojans.dsu.edu

Cherie Noteboom
Dakota State University
cherie.noteboom@dsu.edu

Omar El-Gayar
Dakota State University
omar.el-gayar@dsu.edu

Abstract

This systematic review examines the influence of human cognitive biases on machine learning (ML) systems across the 9 phases of the ML algorithmic value chain. Following the PRISMA guidelines, it synthesizes 19 studies on bias integration and management within ML, highlighting techniques to reduce bias and increase fairness. The review identifies key gaps: the unclear translation of human cognitive biases to ML biases, absence of metrics to measure biases, re-introduction of biases during debiasing, and the critical need for human intervention. These findings prompt several research themes spanning human cognition and algorithmic bias. The theoretical implications are three-fold: extending bias concepts to human cognition, creating an agenda to associate cognitive biases with ML outcomes, and assessing the need for a new or extended discipline. Practically, it raises awareness of human cognition in ML fairness, leading to improved methods for data handling.

Keywords: cognitive bias, fairness, discrimination, machine learning, systematic literature review

1. Introduction

As humans, we harbor inherent values and biases that are imprinted upon us through all our life learned experiences and interactions (McLarney et al., 2021). Unfortunately, we are often unaware of these societal biases and how they, in turn, may frame our future cognitive biases. Cognitive biases are described as inherent tendnecies that cause individuals to deviate from logcal or rational decision-making. These biases arise from the brain's reliance on heuristics—mental shortcuts used to simplify complex information processing, which can lead to systematic errors. While these biases may help in making quick decisions in certain situations, they often result in flawed reasoning, especially in uncertain or complex environments (Korteling & Toet, 2022). Because of this, our very behaviors produce artifacts that exhibit the biases we have (Chouldechova & Roth, 2020). For example, when negoatiating a salary or the price of a car, the first offer acts as an anchor, strongly influencing the follow-up discussions – a phenomena known as anchoring bias. Or, a person that strongly believes in something may seek out information that confirms the bias they already have – known as confirmation bias.

When those biases are captured digitally, as almost every aspect of life is now, they are available for inclusion in systems that use data for any generic decision-making process, up to and including automation of decisions (Mirjam et al., 2021). These biases produce datasets that have qualitative deficiencies that may or may not manifest as quantitative misrepresentations (T. Zhang et al., 2020). Our biases inadvertently produce fundamentally defective data that manifests through defects in quality or quantity that our inherent biases then prevent us from recognizing. Further these biases are then fed back into the datasets that drive future Machine Learning (ML) models leading to systematic digital discrimination (Silva & Kenney, 2019).

Although cognitive bias research has increased in the past twenty years, there is a dearth of research around the application of this research on ML (Fleischmann et al., 2014; Kliegr et al., 2021). To address this gap, this research is a critical examination of the available literature on biases and ML through categorization using cognitive bias category, value chain bias, and debiasing methodology. To motivate this research, we identified the following research questions, each designed to pursue the gaps that exist between the cognitive biases identified by Fleishmann et al. (2014), their location within the algorithmic value chain identified by Silva & Kenney (2019), and their association with mitigation techniques and methods, across the wide range of ML uses.

URI: https://hdl.handle.net/10125/109624
978-0-9981331-8-8

RQ1: What is the association between cognitive biases, the algorithmic value chain, and mitigation techniques?

RQ2: How can these results, within each concept, of cognitive bias, algorithmic value chain, and mitigation techniques be best grouped, organized, or classified?

RQ3: What is the direction of future research necessary to motivate future researchers to address the much-needed gaps in fairness and bias as they are associated with cognitive biases?

The theoretical implications of this research are three-fold: extending bias concepts to human cognition and their associations, developing a research agenda to link cognitive biases to ML outcomes, and determining the need for a new or extended discipline in this area. These implications should drive and inform future research.

Practically, this research raises awareness of human cognition in ML fairness across all disciplines. This awareness could lead to new methods for bias reduction and changes in data collection, use, and dissemination.

2. Background and related work

What is cognitive bias? Cognitive bias is described as a "systematic error in judgement and decision-making common to all human beings which can be due to cognitive limitations, motivational factors, and/or adaptions to natural environments" (Wilke & Mata, 2012). Furthermore, in research, bias can be described as a feature or analysis error that makes evidence misleading (Stegenga, 2018). As identified by Fleischmann at al., in the discipline of Information Systems, there are eight categories of bias: (1) Perception Biases, (2) Pattern Recognition Biases, (3) Memory Biases, (4) Decision Biases, (5) Action-orientated Biases, (6) Stability Biases, (7) Social Biases, (8) Interest Biases. Each of these bias categories has one of more distinct biases (Table 1) that may be applicable to ML (Fleischmann et al., 2014). While the categorization of biases may differ between disciplines, we have adopted Fleischmann et al. (2014).

Category	Biases
Perception biases	framing effect, 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

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

What is fairness in machine learning? Fairness in ML can be divided into two notions, statistical and individual (Chouldechova & Roth, 2020). Most of the previous work on fairness focuses on statistical fairness as it is well defined around bringing parity across known protected demographic groups and equalizes outcomes across non-protected and protected groups (Dwork et al., 2011). The notion of individual fairness brings an attempt to treat similar individuals similarly and attempts to prevent dissimilar individuals from being treated similarly (García-Soriano & Bonchi, 2020). Put simply, similar individuals should be given similar outcomes (Friedler et al., 2021).

What are the phases of machine learning? The PRE phase of ML revolves around the preparation of data and involves all the steps, up to and including selection of training data. It is in this phase that institutionalized, individual, and sample biases are introduced (Bacelar, 2021). The IN phase includes feature selection, algorithm design, and model selection. In this phase assumptions are made and can be generalized as algorithmic biases (Yapo & Weiss, 2018). In the POST phase, model validation is performed, results are synthesized and made available.

There are nine locations where cognitive biases can be introduced into the algorithmic value chain during ML processing, five of which map directly to PRE, IN, and POST-processing phases and are associated with cognitive biases as opposed to the reliance of the output of those biases (Silva & Kenney, 2019): PRE - (1) Training Data Bias, (2) Algorithmic Focus Bias, IN - (3) Algorithmic Processing Bias, POST - (4) Transfer Context Bias, (5) Interpretation Bias, (6) Outcome Non-Transparency Bias, (7) Automation Bias, (8) Consumer Bias, (9) Feedback Loop Bias.

Training Data Bias exists when algorithms are trained on datasets containing our intrinsic biases. It is here that bias first becomes integrated into the ML process, and here that we begin the foundation of potential discrimination on which all additional steps

of ML are built. Algorithmic Focus Bias exists when features are included or excluded from the model. Paradoxically, sometimes those features, such as race or gender must be included to reduce bias, however their inclusion must be weighted appropriately to avoid additional bias (Kamiran & Calders, 2012). Algorithmic Processing Bias is the inclusion of incorrect weighting for features of the algorithm, when performed through supervised learning. Even in unsupervised learning, biases that are not discovered in the training data or feature extraction are potentially included in the model as relevant. Transfer Context Bias exists when non-associated outputs from the ML process are used incorrectly together to draw conclusions, for example if height were used to predict future earning potential. Interpretation Bias exists when the outputs of the models are evaluated and pass through our own internal biases. In all scenarios, self-awareness of bias is paramount, yet something we all struggle to understand within ourselves. Outcome Non-Transparency Bias potentially hides bias within the results. This lack of clarity hinders individuals affected by these decisions from understanding the basis for the outcomes or identifying any discriminatory practices. Automation Bias is simply the trust of the outcome of the process as fact instead of a confidence level. This bias is commonly present in systems that make automated decisions about the outcome of the ML process, such as loan approvals or denials. Consumer bias is the manifestation of the latent biases we have when participating in online activities. Finally, Feedback Loop Bias is created when outputs from ML create data that is then fed back into the same systems and continues to compound that bias into the results.

2.1 "Cognitive Bias Value Chain Framework" (CBVC)

A theoretical and conceptual framework is crucial to understanding and guiding research from development to analysis (Ravitch & Riggan, 2017). By mapping the cognitive biases identified by Fleischmann et al. (2014) to the algorithmic value chain phases proposed by Silva and Kenney (2019), we can better understand the potential impacts of these biases at different phases of the machine learning process. Accordingly, we introduce the "Cognitive Bias Value Chain Framework" (CBVC) to apply these two works together into our theoretical framework. This framework highlights where specific types of cognitive biases are likely to manifest and influence the outcomes of machine learning algorithms. As part of the CBVC, we introduce the "Bias Matrix of Machine Learning" (BMML) model (Figure 1) as the materialization of the CBVC into our conceptual framework, which is designed to systematically categorize and analyze information systems cognitive biases across different phases of the ML process, where they happen, and what the method of bias reduction is. This framework organizes these concepts into four distinct categories: cognitive biases, value chain biases, debiasing methods, and ML processing phases.

The Cognitive Bias Value Chain Framework provides a structured approach to identifying and mitigating biases at every stage of the machine learning process, promoting fair and ethical outcomes. This framework underscores the need for a multifaceted and proactive strategy, leveraging diverse methodologies to ensure transparency, accountability, and equity in machine learning systems.

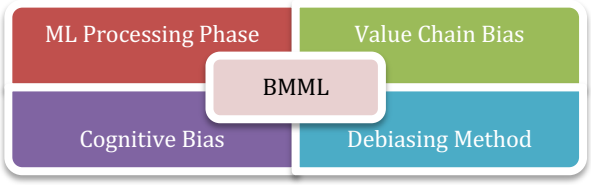

Figure 1. Bias Matrix of Machine Learning

3. Methods

Previous research has highlighted the significance of using an evidence-based method to document and report findings from existing literature (Kitchenham & Charters, 2007). In response to this, our study addresses the research questions by adhering to the systematic literature review (SLR) guidelines for software engineering outlined by Kitchenham & Charters (Kitchenham & Charters, 2007). The main steps involved in carrying out systematic reviews—planning, conducting, and reporting—are detailed below.

The planning phase involves the authors identifying and defining various elements of the research goal, including a preliminary search to ensure the posed questions have not already been addressed in previous reviews. Formulating the research questions introduced earlier was a crucial component of this stage. In developing the search strategy, a preference for title over abstract was employed. Full-text searches were excluded to avoid results that did not focus on the topic of this review. To replicate results fairly across each search engine, search terms were modified to produce result sets that were as similar in content and quality as possible by employing advanced engine-specific search terms.

Following the guidelines set forth by Kitchenham and Charters (Kitchenham & Charters, 2007), a series of search strings derived from the research questions were proposed based on the keywords bias, fairness, inclusion, machine learning, AI, artificial intelligence, human, and cognitive. This list of words captured the totality of concepts and guidelines set forth as the guidelines for this research. Different candidate combinations of these words were used and groupings that produced a high ratio of out-of-scope results were removed from consideration for final usage. Validating each of these combinations and results were performed using title scanning for relevancy.

All searches included bias and fairness interchangeably and required these terms. The term machine learning was also required. Excluded were any results referencing statis* with a wildcard to remove any statistics focused papers which are not within scope of this review. Our search used '(Bias OR Fairness) AND "Machine Learning" AND NOT Statis*' using the following questions as guidelines:

1) Did the paper attempt to identify bias or discrimination specific to machine learning?
2) Did the authors discuss or infer human cognitive bias as part of their research?
3) Was an attempt made to remediate or debias throughout the lifecycle of machine learning?

The review adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines established by Liberati et al., (2009). An extensive search across multiple databases including ACM Digital, IEEE Xplore, ProQuest, Science.gov, and Web of Science was conducted to compile relevant literature. The inclusion criteria allowed for studies from any publication year through April 2024, as long as they remained pertinent to the topic. All studies were from peer-reviewed journals, with pre-print articles only considered if they had been accepted for publication. The search strategy was further refined through manual searches of reference lists in initially identified papers to ensure comprehensive coverage of the subject. To aid in synthesizing the data, an Excel spreadsheet was used to perform coding for theme and sub-themes, with an iterative review and recoding method, establishing the ML processing phase, cognitive bias, value chain bias, and debiasing methods. The following results emerged within the aspects of our CBVC framework.

Each paper chosen for inclusion must address our research questions, namely a cognitive bias was identified within ML, and that bias was successfully mitigated. Crucially, papers focusing on algorithmic bias without a cognitive component were not included, as the focus of this paper is on human introduced, not machine introduced, biases. Papers that merely attempted to categorize bias, understand bias, or apply bias reduction techniques to new domains were also excluded, as they lacked the novel mitigation requirement. The results are organized by machine learning stages and grouped by value chain bias. Explicit biases identified were listed where possible, otherwise bias category was used when necessary.

4. 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 2 for PRISMA flow diagram).

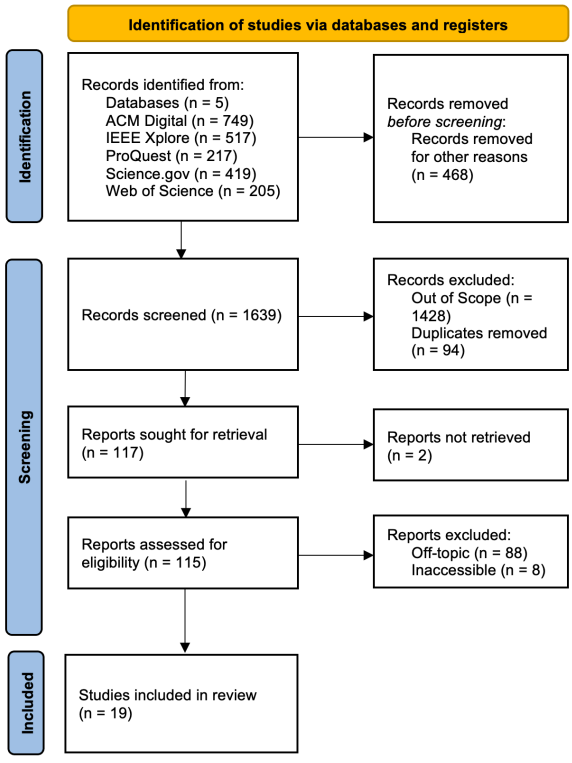

Figure 2. PRISMA Statement

The distribution of the results spanned the published years 2012 to 2023. Overall, the papers meeting the requirements of this SLR are nascent and show an increasing interest, with 2022 having the greatest number of papers with none from 2024.

Results from the systematic review are shown in Table 2. Coding was performed using grouping of methods, key words, and themes and their most likely best fit to each category based on the definition of each concept explored by Bacelar (2021) and Yapo & Weiss (2018).

Interestingly, the grouping of cognitive biases and mitigation exists mostly in the PRE and IN phases of the ML process and aligns with the working processing of building ML models. Since the largest portion of work happens when gathering, cleansing, building, and modeling the data, naturally, the focus will be on these steps when researching remediation issues. Conversely, the smallest grouping of cognitive biases and mitigation happens in the POST phase, which has the most value chain biases.

In addressing *training data bias*, Almuzaini et al. (2022) tackle **stability bias** by preemptively using anticipatory dynamic learning to adjust model parameters, ensuring models remain unbiased despite changing input distributions. Ashokan and Haas (2021) explore **perception biases**, such as observation bias, and propose pre-processing techniques to correct dataset biases before training. Fernando et al. (2021) emphasize the importance of imputing missing values to maintain fairness and address **availability bias**. Johndrow and Lum (2019) remove sensitive information like race from datasets to prevent biased decision-making, tackling **confirmation bias**. Kamiran and Calders (2012) suggest suppressing sensitive attributes and re-weighting to create fair training data, as well as dealing with **confirmation bias**. Wang et al. (2022) and Wang et al. (2023) employ fairness-aware algorithms and incorporate multisource data to train balanced datasets, addressing **social biases** early in the process. Zhang et al. (2020) mitigate decision biases through pseudo-labeling, re-sampling, and ensemble learning to ensure equal representation from the start, targeting **decision biases**.

To address *algorithmic focus bias*, Martin (2019) implements ethical frameworks during development to ensure diverse perspectives, tackling **interest bias**. Nakao et al. (2022) involve end-users in the machine learning process to incorporate human judgment and contextual knowledge, thereby addressing the **framing effect**.

In the IN-processing stage, tackling *algorithmic processing bias* is crucial. Almuzaini et al (2022) address **overconfidence bias** through dynamic and robust learning. Ashokan and Haas (2021) incorporate fairness constraints directly into learning algorithms to address **perception biases**. Ganesh et al. (2023) control data order during training and implement single-epoch adjustments to mitigate **stability bias**. García-Soriano and Bonchi (2020) design fair-by-design matching algorithms to inherently prevent biased outcomes, addressing **representativeness bias**. Ghai and Mueller (2023) use human-in-the-loop approaches and causal modeling to target **overconfidence bias** effectively. Hauptmann et al. (2023) propose maximal representative subsampling to mitigate **selection** and **social biases** by creating representative datasets. Iosifidis et al. (2022) dynamically adjust instance weights during boosting rounds to address **pattern recognition biases**. Taniguchi et al. (2018) incorporate human cognitive biases in their models to enhance learning from small datasets, addressing **pattern recognition biases**. Zhang et al. (2023) use the iFlipper technique to maintain fairness across different groups by flipping labels in real time, addressing the **action-oriented biases** within this ML processing phase.

ML Processing Phase	Value Chain Bias	Cognitive Biases	Sources
PRE	Training Data Bias	Availability, Confirmation, Decision, Perception, Social, Stability	(Almuzaini et al., 2022; Ashokan & Haas, 2021; Fernando et al., 2021; Johndrow & Lum, 2019; Kamiran & Calders, 2012; Wang et al., 2022; Wang, 2023; Zhang et al., 2020)
PRE	Algorithmic Focus Bias	Interest, Framing Effect	(Martin, 2019; Nakao et al., 2022)
IN	Algorithmic Processing Bias	Action-Oriented, Framing Effect, Overconfidence, Pattern Recognition, Perception, Representativeness, Selection, Social, Stability	(Almuzaini et al., 2022; Ashokan & Haas, 2021; Ganesh et al., 2023; García-Soriano & Bonchi, 2020; Ghai & Mueller, 2023; Hauptmann et al., 2023; Iosifidis et al., 2022; Taniguchi et al., 2018; Zhang et al., 2023)
POST	Transfer Context Bias	Framing Effect	(Almuzaini et al., 2022)
POST	Interpretation Bias	Confirmation	(Ghai & Mueller, 2023; Srivastava et al., 2019)
POST	Outcome Non-Transparency Bias	Overconfidence	(Ghai & Mueller, 2023)
POST	Automation Bias	Overconfidence	(Almuzaini et al., 2022)
POST	Consumer Bias	Confirmation	(Martin, 2019)
POST	Feedback Loop Bias	Anchoring	(Almuzaini et al., 2022)

Table 2. Results Chart

Addressing *transfer context bias*, Almuzaini et al. (2022) use dynamic and robust learning to ensure models remain unbiased despite changing input distributions, which helps mitigate the potential for biased outcomes when applying models in new contexts. **Framing Effect** could occur if the model's performance is interpreted differently based on the way outcomes are presented in varying contexts.

In the POST-processing stage, Ghai and Mueller (2023) address *interpretation bias* through interactive visualization tools and causal modeling, helping users identify and mitigate biases effectively. Srivastava et al. (2019) explore how different mathematical definitions of fairness align with human perceptions, using adaptive experiments to reduce cognitive load and enhance fairness interpretations, both addressing **confirmation bias**.

To mitigate *outcome non-transparency bias*, Ghai and Mueller (2023) use causal modeling and interactive visualization tools, helping users understand the decision-making process and outcomes better, thereby reducing the obscurity of the outcomes. **Overconfidence Bias** can occur if users overestimate their understanding and the transparency of the model's decision-making process based on these tools.

Addressing *automation bias*, Almuzaini et al. (2022) use dynamic and robust learning to ensure models remain unbiased over time, maintaining fairness and reliability in automated decision-making processes. **Overconfidence Bias** is relevant here, as users might place undue trust in the automated system's decisions.

In tackling *consumer bias*, Martin (2019) implements ethical frameworks that consider diverse perspectives during the development process, ensuring that algorithmic outcomes do not reinforce existing consumer biases, leading to issues with **confirmation bias**.

Finally, to address *feedback loop bias*, Almuzaini et al. (2022) use anticipatory dynamic learning to adjust model parameters preemptively. This ensures that initial outcomes do not unduly influence future decisions and data, thereby breaking the cycle of biased feedback and limiting **anchoring bias** in future decision-making.

5. Discussion

Exploring cognitive biases and their mitigation within ML systems is crucial for ensuring fair and unbiased outcomes. The results have highlighted various methods proposed in the literature to address cognitive biases across different phases of the algorithmic value chain. The findings underscore the importance of a multi-faceted approach to bias mitigation, incorporating pre-processing techniques, fairness-aware algorithms, dynamic learning strategies, human-in-the-loop systems, and ethical algorithm design, amongst others.

The comprehensive review of the cognitive biases and debiasing methodologies across different stages of the machine learning process provides several key insights:

Multifaceted Nature of Bias: One of the primary insights is the multifaceted nature of bias in machine learning. Bias can occur at multiple stages—from data collection and preprocessing (PRE), through algorithmic processing (IN), to the interpretation and application of results (POST). Each stage introduces different types of biases that require tailored debiasing strategies. This highlights the necessity for a holistic approach to bias mitigation, considering the entire machine learning pipeline rather than isolated components.

Importance of Proactive Bias Mitigation: Proactive strategies, such as anticipatory dynamic learning used by Almuzaini et al. (2022), demonstrate the effectiveness of preemptively adjusting model parameters to address emerging biases. This proactive stance is crucial in maintaining the fairness and accuracy of models over time, particularly in dynamic environments where data distributions can change. By anticipating and mitigating biases before they become entrenched, models can be made more resilient to future changes.

Human-in-the-Loop Approaches: The integration of human-in-the-loop approaches, as seen in Ghai and Mueller (2023), and Nakao (2022), underscores the value of human expertise in identifying and mitigating biases. Human judgment and contextual knowledge can complement algorithmic processes, providing a check against biases that automated systems might overlook. This hybrid approach leverages the strengths of both humans and machines, enhancing overall fairness and accountability. While human intervention could introduce additional cognitive biases, the balance of machine and human, with appropriate awareness, provides a mechanism for applying bounded rationality when appropriate.

Ethical Frameworks and Diverse Perspectives: The emphasis on ethical frameworks and the inclusion of diverse perspectives, highlighted by Martin (2019), is essential for preventing interest biases and ensuring that algorithms do not reinforce existing societal inequities. Ethical guidelines during the development and deployment phases ensure that multiple viewpoints are considered, leading to more equitable outcomes. This aligns with broader societal values and promotes trust in machine learning systems.

Alignment of Mathematical Fairness with Human Perception: Srivastava et al.'s work (2019) on aligning mathematical definitions of fairness with human perceptions points to the need for fairness measures that are both theoretically sound and intuitively acceptable to users. This alignment is crucial for gaining public trust and acceptance of algorithmic decisions. It also suggests that fairness in machine learning is not just a technical challenge but also a social and psychological one.

Importance of Data Handling Techniques: Techniques such as maximal representative subsampling (Hauptmann et al., 2023) and data augmentation (Pastaltzidis et al., 2022) emphasize the critical role of data handling in bias mitigation. Ensuring that training data is representative and balanced is a fundamental step in preventing biases from propagating through the model. These techniques address biases at the source, creating a solid foundation for subsequent algorithmic processes.

Need for Continuous Evaluation and Adaptation: The dynamic nature of machine learning environments necessitates continuous evaluation and adaptation of debiasing strategies. Methods like dynamic and robust learning ensure that models remain fair and accurate over time, adapting to new data and evolving circumstances. This ongoing process of monitoring and adjustment is key to maintaining long-term fairness.

The interconnectedness of Biases: The interconnectedness of different types of biases across the value chain stages indicates that addressing one type of bias may impact others. For example, reducing training data bias can help mitigate algorithmic processing biases and subsequently reduce interpretation biases. This interconnectedness calls for an integrated approach that considers the cascading effects of biases throughout the machine learning pipeline.

These insights underscore the complexity of achieving fairness in machine learning. A multifaceted, proactive, and continuous approach is necessary to address the diverse and interconnected biases that arise throughout the machine learning process. By integrating ethical considerations, leveraging human expertise, and employing robust data handling techniques, the development of more transparent, accountable, and ethical machine learning systems can be realized. This comprehensive strategy is crucial for fostering trust and ensuring equitable outcomes in the deployment of machine learning technologies.

There are four main themes that emerged from this systematic review, categorizable into the four sections of the BMML model (Figure 1). First, no robust research exists for associating human cognitive bias to ML. When developing a process for codifying the results, gaps in current research became apparent as there is a lack of current research that attempts to align human cognitive biases to bias in ML. A more robust coding methodology needs to be built, but current extant research needs to adequately address the concepts necessary to perform this rigorously.

Second, although bias in each stage of ML is well known, the relative values of those biases within each phase and in total are unknown. This lack of quantification in value chain presents an issue with practitioners understanding the return on their bias reduction focus.

Third, that the very act of reducing bias can introduce bias through other human cognitive bias mechanisms, such as applying human-in-the-loop. Extrapolating from the current research on bias and human cognition in the application of bias reduction, the intertwined nature creates a cycle of bias introduction.

Fourth, manual intervention seemed almost inescapable during bias reduction. Following the thread from theme three of bias introduction feedback, manual intervention was simultaneously necessary and a source of a problem.

6. Future Research Agenda

Out of these themes, a proposed research agenda has emerged. This agenda is organized using methodologies incorporated by Jin (2007), presented as a series of probing questions designed to drive a new research direction. Each question is categorized into one of the four identified themes from the discussion.

The questions in the series directly address gaps identified in the systematic review findings and form an attempt to create an extant body of work on this subject as output. The short-term goal of this research is to expand research scope of this systematic review to include additional supportive reviews of crucial concepts. The long-term goal is to provide an extant body of knowledge to build a new research area that addresses a growing and seemingly unknown discipline. The following probing questions will help guide research to fill the knowledge gaps available, and to build a baseline of work that can be expanded in the future.

6.1. Theme 1 – Human Cognitive Bias

Q1-1: What is the association of cognitive bias with machine learning? Although previous works list biases in the Information Systems discipline, the direct

association of those biases to ML is tenuous at best (Fleischmann et al., 2014). Outside of the peer-reviewed works available for systematic review, there were many non-scientific discussions about which cognitive biases were part of the ML landscape. Hardly any of them agree on what cognitive biases are and how to apply them. This represents a significant opportunity for research and a gap in understanding.

Q1-2: Is there a method of tying those cognitive biases to the phases of machine learning that is scientifically sound and will withstand rigor? To answer Q1-1, basic research needs to happen on the methodologies possible to associate cognitive biases with ML. It is possible that this is a novel application of previous research methodologies or even a new field entirely. Too little is known about this subject currently, and significant research will need to be put forth to build rigorous methodologies for the foundation of additional works.

6.2 Theme 2 – Value Chain Identification

Q2-1: Which ML processing phase of the BMML model has the greatest impact on bias reduction? Just because the predominate research exists in Pre-processing does not mean the most impact can be achieved here. A potential avenue of research here would be to determine a framework, if none exists, to evaluate the impact of bias reduction that can be applied to achieve a homogenous score. This score could be used to determine the phases most valuable for bias reduction and, therefore, inform future research avenues. Previous works have attempted to create a measurement of discrimination, so this could be a starting point for determining where to reduce it (Liobait, 2017).

Q2-2: Why is most of the current research in the Pre phase of ML processing? The outcome of the systematic review shows that the grouping of biases is mostly in the PRE-phase of the ML process. Since the largest portion of work happens when gathering, cleansing, building, and modeling the data, naturally, the focus will be on that step when researching remediation issues. However, it's important not to conflate causation and correlation here. We should seek to understand the reasoning behind the predominate research being in this area. Is it motivation, ease of understanding, previous works, or maybe some other phenomena?

6.3 Theme 3 – Introduction of Bias during Bias Reduction

Q3-1: How do practitioners prevent introduction of bias when attempting debiasing methodologies?

Reducing bias and mitigating its effects is a significant endeavor. However, as previous research has shown, bias introduction is intrinsic to humanity (Fuchs, 2018). Since humans are involved in reducing bias, how do we prevent accidentally adding new or different biases into that process? Is it even possible? Researching this question might take the form of categorizing all the available debiasing methods and dividing them into human and non-human groups. Taking these two categories and breaking them down into subgroups would allow specific research into methodologies used and an understanding of where and how potential bias is introduced during remediation. This would aim to enhance understanding of the feedback cycle of bias and inform future work in breaking that cycle.

6.4 Theme 4 – Manual Intervention

Q4-1: Is there a methodology of bias reduction that does not require manual intervention at some level? Not only was the theme of manual intervention consistent with the results of the systematic review, but it was also consistent with previous research performed in this area. There have been two very prominent threads throughout all the work that have led to this research agenda: 1) bias is a critical consideration during all areas of AI, ML included, and 2) the role of the practitioner is critical to the outcome throughout all phases. Given this consistent theme, are there any bias reduction methodologies available that do not depend on manual intervention end-to-end, or will human intervention be required at least at some point? A potential research topic to accomplish this agenda would be to take the debiasing methodologies available at each of the three phases of ML (pre, in, post), aggregate only the ones that can be applied without human intervention, and then try to find direct dependency chains throughout that could potentially uncover an end-to-end methodology.

7. Conclusion

The selected papers employ various debiasing methods targeting different phases of the algorithmic value chain, addressing various cognitive biases. These methods include pre-processing techniques, fairness-aware algorithms, dynamic learning strategies, human-in-the-loop systems, and ethical algorithm design. By implementing these methods, researchers aim to mitigate cognitive biases and ensure fair and unbiased outcomes in ML systems. These debiasing methods provide a comprehensive framework for identifying and mitigating cognitive biases in ML systems, ensuring fair and unbiased

outcomes. Targeting different phases of the algorithmic value chain and addressing specific biases help promote a balanced and ethical approach to ML model development.

The theoretical implications of this research are three-fold: First, it extends the concept of bias to human cognition, its associations, and its co-associated factors. Second, it proposes a research agenda for associating human cognitive biases to ML outcomes that can be used as a foundation for classification. Third, it identifies whether a need exists to extend a current discipline or build a new one in this research area. All these implications should drive new research areas and inform future research in this discipline.

The practical implication is an increased awareness of currently unknown human cognition in fairness in ML throughout its use in all disciplines. This awareness and the resulting new knowledge could drive additional methods for reduction and process changes in collecting, using, and disseminating data.

Concerning the limitations of this review, it should be noted that this review does not attempt to delineate all the issues with bias in ML as it exists in the many different techniques at each step or the specific types of ML methodologies used. Another limitation is the thematization and coding of literature results, as systematic reviews are vulnerable to misclassification errors (Waffenschmidt et al., 2019). Further, in preparation for this paper and in all previous research, we have encountered great difficulty in finding consistency and agreement for many of the concepts around bias in ML.

8. 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. https://doi.org/10.1145/3531146.3533211

Ashokan, A., & Haas, C. (2021). Fairness metrics and bias mitigation strategies for rating predictions. *Information Processing & Management*, *58*(5), 1. ABI/INFORM Collection. https://doi.org/10.1016/j.ipm.2021.102646

Bacelar, M. (2021). *Monitoring bias and fairness in machine learning models: A review* [Preprint]. https://doi.org/10.14293/S2199-1006.1.SOR-.PP59WRH.v1

Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, *63*(5), 82–89. https://doi.org/10.1145/3376898

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

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. https://doi.org/10.1002/int.22415

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. https://doi.org/10.1145/3433949

Fuchs, D. J. (2018). *The Dangers of Human-Like Bias in Machine-Learning Algorithms*. *2*, 15.

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. https://doi.org/10.1145/3593013.3594116

García-Soriano, D., & Bonchi, F. (2020). Fair-by-design matching. *Data Mining and Knowledge Discovery*, *34*(5), 1291–1335. https://doi.org/10.1007/s10618-020-00675-y

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. https://doi.org/10.1109/TVCG.2022.3209484

Hauptmann, T., Fellenz, S., Nathan, L., Tüscher, O., & Kramer, S. (2023). Discriminative machine learning for maximal representative subsampling. *Scientific Reports*, *13*(1), 20925. https://doi.org/10.1038/s41598-023-48177-3

Iosifidis, V., Roy, A., & Ntoutsi, E. (2022). Parity-based cumulative fairness-aware boosting. *Knowledge and Information Systems*, *64*(10), 2737–2770. ABI/INFORM Collection. https://doi.org/10.1007/s10115-022-01723-3

Jin, L., Robey, D., & Boudreau, M.-C. (2007). *A Research Agenda for Investigating Open Source Software User Communities*. 14.

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). https://doi.org/10.1214/18-AOAS1201

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Kitchenham, B. A., & Charters, S. M. (2007). *Guidelines for Performing Systematic Literature Reviews in Software Engineering* (2007–01; ESBE Technical Report). Keele University and Durham University.

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. https://doi.org/10.1016/j.artint.2021.103458

Korteling, J. E. (Hans), & Toet, A. (2022). Cognitive Biases. In *Encyclopedia of Behavioral Neuroscience, 2nd edition* (pp. 610–619). Elsevier. https://doi.org/10.1016/B978-0-12-809324-5.24105-9

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.

Liobait, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, *31*(4), 1060–1089. http://dx.doi.org/10.1007/s10618-017-0506-1

Martin, K. (2019). Designing Ethical Algorithms. *MIS Quarterly Executive*, 129–142. https://doi.org/10.17705/2msqe.00012

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 Reserch Center.

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). http://dx.doi.org/10.1186/s13244-020-00955-7

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. https://doi.org/10.1145/3531146.3534644

Ravitch, S. M., & Riggan, M. (2017). *Reason & rigor how conceptual frameworks guide research* (Second edition). SAGE.

Sengupta, Garg, Choudhury, & Aggarwal. (2018). Techniques to Elimenate Human Bias in Machine Learning. *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*, 226–230. https://doi.org/10.1109/SYSMART.2018.8746946

Silva, S., & Kenney, M. (2019). Algorithms, platforms, and ethnic bias. *Communications of the ACM*, *62*(11), 37–39. https://doi.org/10.1145/3318157

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. https://doi.org/10.1145/3292500.3330664

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. http://dx.doi.org/10.1038/s41598-018-25679-z

Waffenschmidt, S., Knelangen, M., Sieben, W., Bühn, S., & Pieper, D. (2019). Single screening versus conventional double screening for study selection in systematic reviews: A methodological systematic review. *BMC Medical Research Methodology*, *19*(1), 132. https://doi.org/10.1186/s12874-019-0782-0

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. https://doi.org/10.1145/3531146.3533101

Wang, Q., Xu, Z., Chen, Z., Wang, Y., Liu, S., & Qu, H. (2020). Visual Analysis of Discrimination in Machine Learning. *arXiv:2007.15182 [Cs]*. http://arxiv.org/abs/2007.15182

Wang, R. (2023). Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies. *National Library of Medicine's (NLM) PubMed Central (PMC)*. https://doi.org/10.1073/pnas.2211613120

Wilke, A., & Mata, R. (2012). Cognitive Bias. In *Encyclopedia of Human Behavior* (pp. 531–535). Elsevier. https://doi.org/10.1016/B978-0-12-375000-6.00094-X

Yapo, A., & Weiss, J. (2018). *Ethical Implications of Bias in Machine Learning*. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2018.668

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. https://doi.org/10.1109/TKDE.2020.3002567