A step toward building a unified framework for managing AI bias

Saadia Afzal Rana [1], **Zati Hakim Azizul** [Corresp., 1], **Ali Afzal Awan** [2]

[1] Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia

[2] National University of Science and Technology, Islamabad, Pakistan

Corresponding Author: Zati Hakim Azizul
Email address: zati@um.edu.my

Integrating Artificial Intelligence (AI) into every aspect of our experiences has significantly improved living standards. However, AI's efforts are being thwarted by concerns about the rise of biases and unfairness. Conditions become further exacerbated due to multiple interlinked forms of biases and strategies. The problem advocates strongly for the existence of an organized strategy for tackling potential biases. This paper thoroughly evaluates existing knowledge to enhance Bias Management, which will serve as a foundation for creating a unified framework to address any bias and its subsequent mitigation method throughout the AI development pipeline. We map the Software Development Life Cycle (SDLC), Machine Learning Life Cycle (MLLC) and Cross Industry Standard Process for Data Mining (CRISP-DM) process model to have a general understanding of how phases in these development processes are related to each other. The map should benefit researchers from multiple technical backgrounds. Biases are categorized into three distinct classes; Pre-existing, Technical and Emergent Bias, and subsequently, three mitigation strategies; Conceptual, Empirical and Technical, along with Fairness Management Approaches; Fairness Sampling, Learning and Certification. The recommended practices for debias and overcoming challenges encountered further set directions for successfully establishing a unified framework.

A Step Toward Building a Unified Framework for Managing AI Bias

Saadia Afzal Rana[1], Zati Hakim Azizul[1] and Ali Afzal Awan[2]
[1]Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, Universiti Malaya, 50603 Kuala Lumpur, Malaysia
[2]National University of Science and Technology, Islamabad, Pakistan

Corresponding Author:
Zati Hakim Azizul
Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, Universiti Malaya, 50603 Kuala Lumpur, Malaysia
Email address: zati@um.edu.my

ABSTRACT

Integrating Artificial Intelligence (AI) has transformed living standards. However, AI's efforts are being thwarted by concerns about the rise of biases and unfairness. The problem advocates strongly for an strategy for tackling potential biases. This paper thoroughly evaluates existing knowledge to enhance Fairness Management, which will serve as a foundation for creating a unified framework to address any bias and its subsequent mitigation method throughout the AI development pipeline. We map the Software Development Life Cycle (SDLC), Machine Learning Life Cycle (MLLC) and Cross Industry Standard Process for Data Mining (CRISP-DM) together to have a general understanding of how phases in these development processes are related to each other. The map should benefit researchers from multiple technical backgrounds. Biases are categorised into three distinct classes; Pre-existing, Technical and Emergent Bias, and subsequently, three mitigation strategies; Conceptual, Empirical and Technical, along with Fairness Management Approaches; Fairness Sampling, Learning and Certification. The recommended practices for debias and overcoming challenges encountered further set directions for successfully establishing a unified framework.

Subjects Artificial Intelligence, Data Mining and Machine Learning, Data Science, Software Engineering, Emerging Technologies

Keywords Algorithmic Bias; Fairness Management; Bias Mitigation Strategy; Data-Driven AI System; Fairness in Data Mining

INTRODUCTION

. Data-driven decision-making applications have been deployed in vital areas such as finance [1], judiciary [2, 3], employment [4], e-commerce [5], education [6], military intelligence [7] and health [8]. On one side, the potential of AI is widely recognised and appreciated. On the other side, there is significant uncertainty in managing negative consequences and subsequent challenges [9], and a major hindrance to progress is a bias enrooted throughout the AI pipeline's development process [10].

The consequences of unwanted discriminatory and unfair behaviours can be detrimental [11], adversely affecting human rights [12], university admissions [13], profit and revenue [14] and facing legal risks [11,15, 16]. "*Bias has existed since the dawn of society*" [17]. However, AI-based decision-making is criticised for introducing different types of biases. The ever-growing worries demand AI-based systems to rebuild technical strategies to integrate fairness as a core component of its infrastructure. Fairness is the absence of bias or discrimination. An algorithm that makes biased decisions against specific people is considered unfair [18]. The ethical implications of systems that affect individuals' lives have sparked concerns regarding the need for fair and impartial decision-making. Consequently, extensive research has been conducted to address issues of bias and unfairness, while also considering the limitations imposed by corporate policies, legal frameworks, societal norms, and ethical responsibilities [19].

There is still a lack of an organised strategy for tackling potential biases [20]. When looking for the origin of bias in AI decision-making, it is prevalent that the issue is either data or algorithms, and the root cause is humans. Humans transmit cognitive bias while creating/generating data or designing algorithms [9,18]. Thoroughly evaluating existing literature is partial to this work in learning Fairness Management towards proposing a unified framework to address prejudice and its subsequent mitigation method.

In order to gain a comprehensive understanding, we frame our research by mapping the "Software Development Life Cycle" (SDLC), "Machine Learning Life Cycle" (MLLC) and "Cross Industry Standard Process for Data Mining" (CRISP-DM) process model to have a general understanding of how phases in this development process are related to each other. We categorise bias into three classes; Pre-existing, Technical and Emergent Bias and, subsequently, three mitigation strategies; Conceptual, Empirical and Technical, along with Fairness Management Approaches; Fairness Sampling, Learning and Certification.

We discuss them in light of their occurrence at a specific phase of the development cycle. The recommended practices to avoid/mitigate biases and the challenges encountered in the course of action to address them are discussed in the later part.

RATIONALE

A foundation must be established to accommodate all perplexing features, and it is necessary to study various facets of bias and fairness from the perspective of software engineering integration in AI. This study maps together the Software Development Life Cycle (SDLC), Machine Learning Life Cycle (MLLC), and Cross Industry Standard Process for Data Mining (CRISP-DM) processes. The mapping provides a general understanding of how phases in these Software Engineering (SE) and AI development processes relate and can benefit from SE's best practices within AI. The proposed framework aims to detect, identify, and localize biases on the spot and prevent them in the future by comprehending their core causes. The framework handles bias as a defect management process in software engineering.

THE AUDIENCE IT IS INTENDED FOR

We believe we are the first to present this innovative framework. The framework can help ML researchers, ML engineers, data analysts, data scientists, software engineers, and software architects develop superior versions of software applications with higher accuracy, better defect tracking, swifter control test timings and faster time-to-market release.

SURVEY/SEARCH METHODOLOGY SECTION

Integrating bottom-up and top-down research methodologies was used to gather publications on bias in AI/ML applications. Each co-author gathered pertinent material and added it to a shared repository. The keywords for search are optimised in each search with an expectation to shortlist papers containing exact information. The three core domains of artificial intelligence, machine learning, and software engineering were the focus of the search, as shown in Figure 1. Furthermore, 'AND' and 'OR ' string operators were used along with double quotation marks to further narrow down to search fairness in AI/ML.

Interestingly, each survey introduced some new types of bias or fairness definitions. Only papers that detail bias in data, algorithms, assessment tools, fairness management approaches, bias detection, identification,

mitigation strategy, fairness matrices, AI/ML/SE development cycles, datasets characteristics, AI ethics and principles issues were shortlisted during the inclusion criteria. We are further guided to our goal by sections from academic books, keynote addresses by renowned speakers (some of them are J. Buolamwini , T.Gebru and C. O'Neil), and various advisories published. Non-Tabular data, domain/language-dependent technology, and the absence of experimental results are among the exclusion factors that reduce the number of papers we can find from 254 to 72. Due to a lack of mapping and tackling a variety of interlinked biases, we faced many challenges, i.e., uncoordinated development efforts, inefficient use of resources, poor quality of ML models, limited scalability, amplification of biases, unfair decision-making, a lack of diversity, a loss of trust, and ethical issues. And in short found difficulties in integrating ML models into software. Addressing these issues requires a multifaceted approach that involves addressing biases at multiple levels, including in the data, the algorithms and the decision-making processes. It also requires a commitment to ethical and responsible development and use of technologies, especially when all development cycles should be synchronised together.

REVIEW OF LITERATURE

The literature review on the subject is dense, with many forms of biases and mitigation strategies, most of which are interlinked. It is essential to be aware of potential biases in the AI pipeline and the appropriate mitigation strategies to address undesirable effects [21].

Realistic Case Study of AI Bias

The previous ten years have seen AI's widespread adoption and popularity, allowing it to permeate practically every aspect of our daily lives. However, safety and fairness concerns have prompted practitioners to prioritise them while designing and engineering AI-based applications [22, 23, 24]. Researchers have enumerated a few applications that, because of biases, have a detrimental impact on people's lives, such as biometrics apps, autonomous vehicles, AI chat-bots, robotic systems, employment matching, medical aid, and systems for children's welfare.

The US judiciary uses the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) [34], an AI-based tool, to identify offenders more likely to commit crimes again. However, the Pulitzer Prize-winning nonprofit news organisation ProPublica [2] discovered that COMPAS was racially

prejudiced and that black criminals were at high risk [25]. PredPol, or predictive policing, is an AI-enabled software to predict the next crime location based on the number of arrests and frequency of calls to the police regarding different offences. PredPol was criticised for its prejudiced behaviour as it targeted racial minorities [26].

An AI algorithm known as the Amazon recruiting engine was developed to evaluate the resumes of job applicants applying to Amazon and then shortlist eligible candidates for further interviews and consideration. However, it turned out that the Amazon algorithm was discriminatory regarding hiring women [27]. A labelling tool in Google Photos adds a label to a photo that corresponds to whatever is seen in the image. When it referred to images of a black software developer and his friend as gorillas, it was determined to be racist [28]. The well-known programme StyleGAN [29] automatically creates eerily realistic human faces and produces white faces more frequently than racial minorities.

Biases based on racial, gender, and other demographic factors restrict communities from using AI-automated technologies for regular tasks in the health and education sectors [30, 31]. Organisations must know the different types of biases in their data/algorithm that can affect their Machine Learning (ML) models. Ultimately, identifying and mitigating biases that skewed or produced undesirable outcomes and impeded the advancements made by AI for the everyday person is more than necessary.

Assessment Tools

A systematic effort has been made to provide appropriate tools for practitioners to adopt cutting-edge fairness strategies into their AI pipelines. Software toolkits and checklists are the two basic ways to ensure fairness. Programming language functions that can be used to identify or lessen biases are known as toolkits. AI practitioners can employ checklists and detailed instructions by fairness specialists to ensure ethical consideration is incorporated across their pipelines [21].

Fairness Indicators and What-If toolkit (WIT) are well-known tools that Google provides [20]. Fairness Indicators are based on fairness metrics for binary and multiclass classifiers. What-If Tool is an interactive visual tool designed to examine, evaluate, and compare ML models [20]. Uchicago's Aequitas is utilised to assess ML-based outcomes to identify various biases and make justified choices regarding the creation and implementation of such systems [32]. IBM's AI Fairness 360 is a toolkit providing fairness detection

and mitigation strategies [33]. LinkedIn's Fairness Toolkit (LiFT) provides detection strategies for measuring fairness across various metrics [31]. Microsoft's Fairlearn, ML Fairness Gym, Scikit's fairness tool, and PyMetrics Audit-AI are readily available tools to detect/mitigate (or both) bias and ensure fairness management. These assessment techniques expand practitioners' options for developing ethical products and maintaining stakeholder confidence.

Despite the availability of tools that incorporate explainable machine learning methods and fair algorithms, none of them currently offers a comprehensive set of guidelines to assist users in effectively addressing the diverse fairness concerns that arise at different stages of the machine learning decision pipeline[34].Moreover the responsibility for identifying and mitigating bias and unfairness is often entirely placed on the developer, who may not possess sufficient expertise in addressing these challenges and cannot be solely responsible for the same [35] . There is lack of consensus on what constitutes fairness as there is no single definition of fairness. Moreover, these tools have limited ability to detect complex forms of bias as they are designed to detect bias due to protected attributes only and casual fairness demands to investigate underlying relations between different attributes. Fairness assessment tools can themselves be biased. For example, a tool that is designed to detect bias in natural language processing models may be biased towards certain languages or dialects. Sometimes they can be time-consuming and expensive to use for organizations to adopt these tools, especially small or resource-constrained organizations. One more notable objection regarding them is that it can be difficult to interpret the results which make it difficult for organization to take action to address any bias that is found. Majority of fairness assessment tools may not be able to identify the root cause of bias.

Fairness Management Research Datasets

Several standard datasets are available to make research on bias and fairness easier. Each of them has sensitive or protected attributes that could be used to demonstrate unfair or biased treatment toward underprivileged groups or classes. Table 1 summarises some of these datasets' characteristics. These datasets are used as a benchmark to evaluate and compare the effectiveness of bias detection and mitigation strategies.

BROAD SPECTRUM OF BIAS

Since bias has existed for as long as human civilisation, research on the topic is popular. The literature review, however, is full of various terminologies and theories that either overlap or are interlinked, and this

conflict is enough to perplex researchers. In this section, we will try to throw light on a different aspect of Bias in AI with a perspective to overcome confusion and increase researchers' understanding. This effort sets the groundwork for building a unified framework for Fairness Management.

Bias, Discrimination and Unfairness

The existence of bias, discrimination, and unfairness are related topics, but it is essential to distinguish them for further investigation/analysis [44]. "*Bias is the unfair inclination or prejudice in a judgement made by an AI system, either in favour of or against an individual or group."* [18]. Discrimination can be considered a source of unfairness due to human prejudice and stereotyping based on sensitive attributes, which may happen intentionally or unintentionally. In contrast, bias can be considered a source of unfairness due to the data collection, sampling, and measurement [45]. Discrimination is a difference in the treatment of individuals based on their membership in a group [46]. Bias is a systematic difference in treating particular objects, people or groups compared to others. Unfairness is the presence of bias where we believe there should be no systematic difference [4]. Bias is a statistical property, whereas fairness is generally an ethical issue.

Protected, Sensitive and Potentially Biased Attributes

Until we stop making it, machine bias will keep appearing everywhere: bias in, bias out [47]. Protected attributes are characteristics that cannot be relied upon to make decisions and may be chosen based on the organisation's objectives or regulatory requirements. Similarly, sensitive attributes are characteristics of humans that may be given special consideration for social, ethical, legal or personal reasons [48]. In literature, these terms are used interchangeably and in place of one another. It is a potentially biased attribute if changing the value of an attribute through the alternation function has an impact on prediction, such as altering a male attribute to a female one or a black attribute to a white one [47]. Any one of these attributes in the dataset necessitates special consideration. Sex, Race, Color, Age, Marital Status, Family, Religion, Sexual Orientation, Political Opinion, Pregnancy, Physical or Mental Disability, Career Responsibilities, Social Origins, and National Extraction are a few attributes examples [49]. However, it has been observed that groups/individuals still face discrimination through proxy attributes, even in the absence

of some protected or sensitive traits. Proxy attributes correlate with protected or sensitive attributes, such as Zip code (linked with Race) [12].

Trio-Bias Feedback Loop

Data is the primary driving force behind most AI systems and algorithms, so they need data to be trained. As a result, the functioning of these algorithms and systems is closely tied to data availability. Underlying training data bias will manifest through an algorithm's predictions (trained on it). Furthermore, algorithms may reflect discriminatory behaviour based on specific design considerations even when the data is fair. The biased algorithm's output can then be incorporated into the existing system and impact users' choices, producing more biased data that can be used to train the new algorithm. Figure 2 depicts the feedback loop between data biases, algorithmic biases, and user involvement. Humans are involved in data preparation and algorithm design; therefore, whether bias results from data or an algorithm, humans are the root cause.

Categorisation of Biases

The human bias has more than 180 varieties, for example 50]. Bias has been broadly divided into three major categories to accomplish fairness with a focus on detection and mitigation mechanisms and dispel confusion with many different types of bias [51]. Bias categories include Pre-Existing, Technical and Emergent [52].

A bias before the development of technology is referred to as Pre-Existing Bias. This kind of bias has its roots in social structures and manifests itself in individual biases. The same is introduced into technology by people and organisations responsible for its development, whether explicitly or implicitly, consciously or unconsciously [52]. The AI literature identifies that the most typical biases are pre-existing [51]. A few examples of Pre-Existing Bias are the wrong model of Microsoft bot [53], Duane Buck's murder [54], the stance on fairness [54], criminal justice models [55], and the understanding of concepts and reasonability of CO2 emissions [56].

Technical bias refers to concerns about a product's technological design, such as technical limitations or decisions [52]. Technical issues faced in prominent software like IMPACT, Tech, LSI-R, Kyle's job application and hiring algorithms are because of Technical Bias. Emergent bias emerges after the

practical use of a design as a result of a shift in social awareness or cultural norms [52]. PredPol, St. George's model, COMPAS, and facial analysis tech were adversely affected due to Emergent Bias.

Origins of Biases within the Development Cycle

The literature regarding the causes of biases and potential mitigation strategies is still scattered, and work on a systematic methodology for dealing with potential biases [57] and building well-established ML frameworks is in progress [, 58, 59, 60]. In this paper, we attempt to overcome confusion and enhance understanding regarding the origin of different categories of bias during the development cycle by mapping SDLC, MLLC and CRISP-DM on a single scale, as shown in Figure 3.

Professional designers or developers introduce Pre-Existing Bias into the technical process, which is why it drags into the modelling phase of the CRISP-DM cycle [61]. Technical bias, as opposed to Pre-Existing Bias, results from how problems in the technical design are resolved. The design process and model evaluation contain many areas where Technical Bias can be identified [52, 63]. Pre-Existing and Technical Biases occur prior to and within the technical development process. However, Emergent Bias appears during the actual use of the technical product after development [61]. Emergent bias can be detected before deployment (during testing) and is often the most obvious category of bias [51]. The same type of bias has several names in the literature. Therefore, biases were descriptively synthesised and characterised based on their origins [63, 64]. Based on a thorough understanding of the development cycles, commonly encountered distinct types of biases were assigned to the phases based on their origin, as shown in Figure 4.

The bias type and subtype distribution across different phases will create an appropriate detection method [10]. Equal Opportunity, Equalized Odds, Conditional Demographic Disparity, Disparate Impact, Euclidean distance, Mahalanobis distance, Manhattan distance, Demographic Disparity, and different tools and libraries are used for bias detection [65].

Bias Minimising Strategies

Value Sensitive Design (VSD) is a methodology that can contribute to understanding and addressing issues of bias in AI systems[67] and to promote transparency and ethical principles in AI systems[68]. .It is a framework that provides recommendations for minimising or coping with various biases [51,52], is adopted in this research

as a potential strategy to minimise the biases associated with AI. In the VSD study, “value” is a general phrase that relates to what user values in life [62]. This theoretically-based approach provides three possible solutions based on the types of investigations to prevent problems and advocate AI-specific value-oriented metrics that all stakeholders mutually agreed on.

Conceptual, Empirical and Technical are three different kinds of investigations. A conceptual investigation primarily focuses on analysing or prioritising different stakeholders’ values in the design and use of technology [61, 66]. Empirical investigation assesses a technical design’s effectiveness using factors like how people react to technological products [61, 66]. It frequently entails observation and recording; quantitative and qualitative research methods are appropriate. Because of these characteristics, the phases are limited to Data Understanding (slightly), Data Preparation, modeling (to some extent), Deployment and Feedback.

The primary focus of the technical investigation is on technology while focusing on how technological characteristics and underlying mechanics promote or undermine human values and involves proactive system design to uphold values discovered in conceptual investigations [51, 52, 63]. VSD has revealed the relationships between different types of investigations and types of AI biases. Researchers concluded by recognising the value of conceptual and empirical investigation for addressing Pre-Existing Bias, investigations for minimising Technical Bias, and technical and empirical investigation for tackling Emerging Bias [51].

Bias Mitigation Methods

Several methods can mitigate a single bias, and multiple biases can be mitigated by a single method. Socio-technical approach comprising technical and nontechnical methods is widely adopted to counterattack the harmful effects of biased decisions [10]. This approach mitigates bias and prevents it from recurring in the future. Table 2 depicts which method effectively mitigates different types of bias whenever it appears in different stages of the AI development cycle. The mitigation process is not executed during Feedback. From Risk Management’s perspective, to evaluate an AI system’s performance, ‘Fairness’ outclass any other vital measure, i.e., dependability, efficiency, and accuracy [61].

Various strategies have been put forth by researchers and practitioners to address bias in AI. These strategies encompass data pre-processing, model selection, and post-processing decisions. However, each of these methods has its own set of limitations and difficulties, such as the scarcity of diverse and representative training data, the complexity of identifying and quantifying different forms of bias and the potential trade-offs between fairness and accuracy [34]. Additionally, ethical considerations arise when determining which types of bias to prioritize and which groups should be given priority in the mitigation process [69]. Bias mitigation methods can be complex , computationally expensive and can introduce new biases for example, a method that tries to balance the representation of different groups in a dataset may introduce a new bias in favor of the majority group. They can be brittle such as that they can be sensitive to changes in the data or the model. This can make it difficult to ensure that the model remains fair over time. Another notable aspect that needs to be addressed are that there is a lack of understanding of the long-term effects of bias mitigation methods & standardized evaluation metrics. These methods can be opaque as it can be difficult to understand how they work ,subsequently it can make it difficult to trust these methods, and to ensure that they are not introducing new biases. They might feel it difficult to adapt to new tasks, as result of it , they may not be effective for all machine learning models.

Despite these obstacles, the mitigation of bias in AI is of utmost importance to establish just and equitable systems that benefit everyone in society [70]. Continuous research and development of mitigation techniques are crucial to overcome these challenges and ensure that AI systems are employed for the welfare of all individuals [71].

FAIRNESS MANAGEMENT APPROACH

Bias and fairness are two mutually exclusive aspects of reality. In the absence of a unified definition, an absence of bias or preference for individuals or groups based on their characteristics is generally regarded as ‘Fairness’. It is necessary to do more than mitigate any bias detected to ensure that an AI system may be regarded as fair. Instead, “*fairness-aware*” system design should be encouraged [72]. This approach incorporates “fairness” as a crucial design component from conception to deployment (sometimes extended to maintenance or upgradation). Figure 5 shows the fairness management approach and several implementation methods across the CRISP-DM model. The following steps constitute the fairness management approach:

Fairness Formalisation

Constraints, measures, specifications and criteria to ensure fairness is defined during the business understanding and data understanding phases [15], as depicted in Figure 5. In fact, at this stage, the benchmarks for auditing or evaluating fairness are stated.

Fairness Sampling

Fairness sampling generally refers to preprocessing skewed data through different methods, such as oversampling [72]. Issues with data, i.e., inaccuracy, incompleteness, improperly labelling, too much/less, inconsistency, and silos, are addressed at this stage [73]. Fair sampling is accomplished during the data understanding and data preparation phases.

Fairness Learning

It is not always feasible to develop a fair model by eliminating the bias in the initial data before training. Designing a fair classifier that uses a fair algorithm is the solution in such scenarios [74]. As a result, we can still use a biased dataset to train the model, and the fair algorithm still produces predictions through in-processing methods carried out during the modelling and evaluation phases [74].

Fairness Certification

At the final stage of the testing and deployment phases, it is evaluated whether a prediction aligns with the criteria mentioned in the fairness formalisation phase by executing post-processing methods [66]. Fairness certification solutions verify that unfairness does not surface during the feedback phase.

Limitations and Challenges of Existing Approaches

After thoroughly examining various aspects of tools, methodologies, frameworks, and fairness solution spaces, it is now appropriate to consolidate and summarize the overall major limitations and challenges encountered throughout this exploration:

1. One of the shortcomings of current approaches is the lack of transparency and interpretability [75]. Many tools and strategies employed to mitigate biases in various domains, such as machine learning algorithms or content moderation systems, often lack clear explanations of how they address bias. This lack of transparency makes it difficult for users to understand the underlying

biases being addressed and the effectiveness of the applied methods. Additionally, without proper transparency, it becomes challenging to identify potential unintended consequences or biases that may arise from the bias management methods themselves.

2. Another major shortcoming lies in the limited customization and adaptability [76]. Most tools and strategies are developed with a one-size-fits-all approach, which may not adequately account for the specific biases prevalent in different contexts or domains[77]. This limits their effectiveness in managing biases that are nuanced and context-dependent. Furthermore, these methods often do not provide enough flexibility for users to customize and fine-tune the bias management mechanisms according to their specific needs and requirements.
3. Existing methods, tools, and strategies for bias management rely on manual intervention, which makes the process time-consuming and prone to human error. When bias management is carried out manually, it becomes difficult to ensure consistent and comprehensive coverage of all potential biases. Additionally, manual methods may lack scalability and efficiency, particularly when dealing with large datasets or complex models[78]. Therefore, there is a need for more automated and robust approaches to bias management.
4. Bias can be unintentionally introduced when the dataset used to train AI models is not representative of the real-world population it is designed to serve[79]. For example, if a facial recognition system is trained primarily on data from lighter-skinned individuals, it may exhibit higher error rates for darker-skinned individuals. This lack of diversity can perpetuate existing societal biases and lead to discriminatory outcomes.
5. Existing methodologies tend to place a heavy emphasis on technical solutions for bias mitigation, often neglecting the importance of interdisciplinary collaboration [80]. Addressing bias in AI requires input from diverse stakeholders, including ethicists, social scientists, and policymakers. Without incorporating a multidisciplinary perspective, frameworks may overlook crucial ethical considerations and fail to account for the broader societal impact of AI systems.
6. Many existing strategies tend to oversimplify the concept of bias, reducing it to a binary problem. They often focus on mitigating only explicit biases while overlooking implicit biases, which are more subtle and deeply ingrained in societal structures [81]. Addressing implicit biases requires a

more nuanced understanding of the underlying social dynamics and power structures. Failure to consider these complexities can result in incomplete or ineffective bias mitigation strategies.

7. Identifying and mitigating Identifying and mitigating the various interlinked biases that can arise in AI systems poses a significant challenge due to their diverse nature and complexity.
8. In some cases, there may be a potential trade-off between fairness and accuracy. For example, if an AI system is designed to be fair to all groups, it may not be as accurate as it could be.
9. There are ethical considerations around how to mitigate bias in AI systems. For example, should we prioritize fairness to individuals or to groups? Should we focus on mitigating historical bias or on preventing future bias.
10. The prevailing focus of most approaches lies in addressing bias reactively, leading to high costs associated with corrective measures. There is a pressing need to adopt a proactive stance and mitigate bias as soon as it is identified

Lack of comprehensive evaluation frameworks is another hurdle in way to achieve fairness. While some methods may claim to address biases, there is often a lack of standardized evaluation frameworks to assess their effectiveness and potential trade-offs [82]. This absence of robust evaluation frameworks hinders the attainment of fairness on a global scale.

RECOMMENDED PRACTICES TO AVOID/MITIGATE BIAS

After understanding the categorisation and minimising strategies of biases and knowing how to implement a fairness management approach during the model development life cycle, it is time to look at best practices to avoid/mitigate bias and ensure fairness. A few of them are as follows:

- Considering human customs in AI use and promoting all concerned stakeholders on the board i.e developers, users /general public, policy makers etc. to achieve an effective strategy [66].
- Domain-specific knowledge must be incorporated to detect and mitigate bias [83].
    - When collecting data, it is vital to have expertise in extracting the most valuable data variables [84].

    - Be conscious of the data's sensitive features, including proxy features, as determined by the application.
    - Datasets should, to the greatest extent possible, represent the actual population being taken into account. Data selection by random sampling can perform effectively [84].
    - Preprocess the data to guarantee the maximum possible level of accuracy while minimising relations between results and sensitive attributes or to present data preserving privacy [66].
    - For annotating the data, appropriate standards must be specified.
    - Consider crucial elements such as the data type, problem, desired outcome, and data size while choosing the best model for the data set [84].
- The right model choice is one of the critical components of fairness management. Compared to linear models, which give exact weights for each feature being taken into account, deep models like decision trees can more easily conceal their biases [85].
- Bias detection should be incorporated as a necessary component of model evaluation and focus on model accuracy and precision [86].
- Track the models' performance in use and continually evaluate it [86].
- Encourage all stakeholders to report bias and management should take it positively [87].
- In order to prevent perpetuating inequity, AI systems must be responsible during the design, development, evaluation, implementation, and monitoring phases [88].
- Process and result transparency should be defined in a way that can be interpretable without in-depth knowledge of the algorithm [89].

CHALLENGES, OPPORTUNITIES AND FUTURE WORK

There are several obstacles to overcome before ethical AI applications and systems are successfully developed, free from bias, and wholly engineered along fairness lines. Planning to overcome these obstacles sets the direction of future research. A few serious challenges are:

- Despite numerous approaches to detect and mitigate bias /unfairness, no absolute results are yet available for the cutting edge to handle each type of biasness [90,34].

- A mathematical definition cannot express all notions of fairness [90]. From the ML perspective, literature is enriched with various definitions of fairness. One of the unsolved research issues is how to combine these definitions and propose a unique notion of fairness [91]. Achieving it can evaluate AI systems in more unified and comparable manners. The ideals of fairness are incompatible with the operationalisation of the “from Equality to Equity” concept [18], which necessitates approaching the issue from an operational war posture. Moreover, it demands integrating social and political knowledge in the primary process as critical elements [92].
- The literature review has a lot to say about the bias/fairness of the data and algorithm used by data-driven decision-making systems. However, not all areas have received the same degree of research community attention, i.e., classification, clustering, word embedding, semantic role labelling, representation learning VAE, regression, PCA, named entity recognition, machine translation, language model, graph embedding, coreference resolution, and community detection [12].
- In the context of fairness-aware ML, exploratory analysis of datasets is still not practiced widely [91]. Real, synthetic and sequential decision-making datasets are not adequately exploited.
- The role of sensitive /protected attributes in measuring the performance of predictive models has been studied a lot. However, the role of proxy attributes in the same perspective demands more research work [91].
- Most research being done at the moment focuses on techniques that reduce bias in the underlying machine learning models through the algorithm’s approach. Due to this, a research gap may be filled by a data-centered approach to the subject [92].
- The oldest benchmark dataset was gathered 48 years ago from nations with active data protection laws. However, to meet the demands of the modern day, general data quality or collection regulations still need to be researched and developed [17].

Establishing an Agile Approach to Address Bias in AI Systems:

A framework for managing bias in AI systems should be created to encourage fairness, accountability, and transparency throughout the AI system lifetime while integrating software engineering best practices, in

light of the challenges listed above. Effectively reducing bias requires a multifaceted, adaptable, transparent, scalable, accessible, interdisciplinary, and iterative approach.

Agile methods can be employed to address bias in AI systems. By adopting an agile approach, AI developers can continuously monitor and mitigate bias throughout the development lifecycle and all stakeholders will be well informed with current situation [94]. Framework based on agile approach can not only mitigate bias at the spot in proactive manners but also eradicate it from appearing in future by eliminating all interlinked biases as [82 , 95]. To design a framework for fair AI several working variables play a key role. Let's explore some of these variables in detail:-

- **Data Collection and Preprocessing**

The first working variable to consider is the data used to train AI models. It is essential to ensure that the data collected is representative and diverse, without any inherent biases [97]. Biases can emerge if the data reflects historical prejudices or imbalances [98]. Careful preprocessing is necessary to identify and address these biases to prevent unfair outcomes. For example, if a facial recognition system is trained predominantly on a specific demographic, it may exhibit racial or gender biases.

- **Algorithmic Transparency and Explainability**

To design a fair AI framework, it is crucial to consider the transparency and explainability of the algorithms used. Black-box algorithms that provide no insight into their decision-making processes can pose challenges in identifying and rectifying biases [98]. By promoting algorithmic transparency, stakeholders can understand how decisions are made and detect any unfairness in the system. Explainable AI techniques, such as providing interpretable explanations for decisions, can also enhance fairness and accountability [96].

- **Evaluation Metrics**

Establishing appropriate evaluation metrics is another essential working variable in designing a fair AI framework. The metrics used to assess the performance of AI systems should go beyond traditional accuracy measures and incorporate fairness considerations [98]. For instance, metrics like disparate impact, equal opportunity, and predictive parity can help identify and mitigate biases across different demographic groups.

Evaluating AI systems on these fairness metrics ensures that fairness is a fundamental goal rather than an afterthought.

- **Regular Audits and Monitoring**

A fair AI framework requires ongoing audits and monitoring to identify and rectify biases that may emerge over time. Regular assessments can help evaluate the fairness of AI algorithms and models in real-world scenarios [82]. It allows for continuous improvement and ensures that any biases are promptly addressed. Organizations should establish mechanisms to monitor the performance of AI systems, collect feedback from users, and engage in iterative improvements to enhance fairness. Moreover, encourage continuous learning within your team about bias, fairness, and ethical AI. Continuously stay informed about the most recent research and best practices in this domain and ensure ongoing updates.

- **Stakeholder Inclusivity**

Involving diverse stakeholders in the design and deployment of AI systems is a critical working variable for fair AI. Including representatives from different communities, demographic groups, and experts from various fields can help identify potential biases and ensure fairness [98]. It is crucial to consider the perspectives and experiences of those who may be disproportionately affected by AI systems. Such inclusivity can lead to a more comprehensive understanding of biases and result in fairer outcomes.

- **Ethical Guidelines and Governance**

Ethical guidelines and governance play a vital role in shaping the design of a fair AI framework. Organizations should establish clear policies and guidelines that explicitly address fairness concerns. These guidelines should define what constitutes fairness and provide actionable steps to ensure it is upheld. Integrating ethical considerations into the design and decision-making processes can help prevent biases[99].

In nutshell, agile methods provide a framework for effectively addressing bias in AI systems. By adopting diverse teams, defining clear goals and metrics, conducting frequent reviews and iterations, ensuring transparency, curating unbiased datasets, monitoring and evaluating system performance, collaborating with stakeholders, and prioritizing ethics, organizations can develop and deploy AI systems that are fair, unbiased, and inclusive.

CONCLUSION

The presence of bias in our world is reflected in the data. It can appear at any phase of the AI model development cycle. It's not only developers' core responsibility is to ensure model fairness but AI fairness requires a collaborative effort from all stakeholders i.e policymakers, regulators, users, general public etc. for responsible and ethical AI solutions[92]. The future will see AI play an even more significant impact in both our personal and professional lives. Recognising the benefits and challenges of developing ethical and efficient AI models is crucial. AI developers bring with them a variety of disciplines and professional experiences.

Boring the subject to a single profession or area of expertise would oversimplify the situation. Mapping SDLC, MLLC, and CRISP-DM on a single reference will boast an understanding of practitioners from various technical frameworks to a single point. Once the bias is identified and mitigated at each phase of the development process, the technical team will remain vigilant and unable to de-track or display negligence. The fairness management approach will further enhance the effectiveness of revealing the hidden shortcomings during the development process.

From this perspective, a firm grasp of standard practices will be the foundation for a unified AI framework for fairness management. Through the proposed framework, organisations of all sizes can manage the risk of bias throughout a system's lifecycle and ensure that AI is accountable by design. In order to manage the associated risks with AI bias, the proposed framework ought to have the following key characteristics:

- Outlines a technique for conducting Impact Assessments
- Promotes better awareness of already-existing standards, guidelines, recommendations, best practices, methodologies, and tools and indicates the need for more effective resources.
- Integrate best practices from Software Engineering similar to 'defect management' to tackle bias as 'defect'. Then, detect, identify and localise bias on the spot before proceeding further and eliminating its chance of appearing in future.
- Ensure each stakeholder comprehends his or her responsibility.

- Sets out corporate governance structures, processes and safeguards that are needed to achieve desired goals
- Law and regulation agnostic
- In light of the field of AI's rapid growth, the framework should be updated to ensure it is up-to-date and adapted

Briefly, AI fairness management is centered on a governance framework that encourages the prevention of bias from manifesting in a way that unjustifiably leads to less favorable or harmful outcomes and enables businesses to create more accurate and practical applications and more persuasive to customers. Overall, a framework for fair AI should provide a structured approach to manage biases and ensure that AI systems operate ethically, fairly, and transparently, while accommodating the complexities and challenges inherent in AI development and deployment.

REFERENCES

[1] Baum, Seth D. "On the promotion of safe and socially beneficial artificial intelligence." *AI & SOCIETY* 32.4 (2017): 543-551.

[2] Angwin, J., Larson, J., Mattu, S., Kirchner, L., Baker, E., Goldstein, A. P., & Azevedo, I. M. (2016). Machine Bias. Ethics of Data and Analytics. *Energy and Climate Change, 2*.

[3] Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., & Robinson, D. G. (2020, January). Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 252-260).

[4] Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K. W. (2018). Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.

[5] Weith, H. V. K., & Matt, C. (2022). When Do Customers Perceive Artificial Intelligence as Fair? An Assessment of AI-based B2C E-Commerce.

[6] Kim, J., Lee, H., & Cho, Y. H. (2022). Learning design to support student-AI collaboration: perspectives of leading teachers for AI in education. *Education and Information Technologies*, 1-36.

[7] Maathuis, C. (2022, March). On Explainable AI Solutions for Targeting in Cyber Military Operations. In *International Conference on Cyber Warfare and Security* (Vol. 17, No. 1, pp. 166-175).

[8] Gardner, A., Smith, A. L., Steventon, A., Coughlan, E., & Oldfield, M. (2022). Ethical funding for trustworthy AI: proposals to address the responsibilities of funders to ensure that projects adhere to trustworthy AI practice. *AI and Ethics*, *2*(2), 277-291.

[9] There's more to AI bias than biased data, NIST report highlights. (2022, March 16). Retrieved from https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights

[10] Fahse, T., Huber, V., & Giffen, B. V. (2021, March). Managing bias in machine learning projects. In *International Conference on Wirtschaftsinformatik* (pp. 94-109). Springer, Cham.

[11] Pedreshi, D., Ruggieri, S., & Turini, F. (2008, August). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 560-568).

[12] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35.

[13] Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex Bias in Graduate Admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, *187*(4175), 398-404.

[14] Mikians, J., Gyarmati, L., Erramilli, V., & Laoutaris, N. (2012, October). Detecting price and search discrimination on the internet. In *Proceedings of the 11th ACM workshop on hot topics in networks* (pp. 79-84).

[15] News@Northeastern. (2020, April 23). Here's what happened when Boston tried to assign students good schools close to home. Retrieved from https://news.northeastern.edu/2018/07/16/heres-what-happened-when-boston-tried-to-assign-students-good-schools-close-to-home/

[16] Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, *29*(5), 582-638.

[17] Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1356.

[18] Fenwick, A., & Molnar, G. (2022). The importance of humanising AI: using a behavioral lens to bridge the gaps between humans and machines. *Discover Artificial Intelligence*, *2*(1), 1-12.

[19] Hobson, Z., Yesberg, J. A., Bradford, B., & Jackson, J. (2021). Artificial fairness? Trust in algorithmic police decision-making. *Journal of experimental criminology*, 1-25.

[20] Richardson, B., Garcia-Gathright, J., Way, S. F., Thom, J., & Cramer, H. (2021, May). Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

[21] Bailey, D., Faraj, S., Hinds, P., von Krogh, G., & Leonardi, P. (2019). Special issue of organisation science: Emerging technologies and organising. *Organization Science*, *30*(3), 642-646.

[22] Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018, January). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency* (pp. 134-148). PMLR.

[23] Howard, A., & Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and engineering ethics*, *24*(5), 1521-1536.

[24] Osoba, O. A., & Welser IV, W. (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation.

[25] Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, *4*(1), eaao5580.

[26] Benbouzid, B. (2018). Values and Consequences in Predictive Machine Evaluation. A Sociology of Predictive Policing. *Science & Technology Studies*, *31*.

[27] Hofeditz, L., Mirbabaie, M., Luther, A., Mauth, R., & Rentemeister, I. (2022, January). Ethics Guidelines for Using AI-based Algorithms in Recruiting: Learnings from a Systematic Literature Review. In *HICSS* (pp. 1-10).

[28] González Esteban, E., & Calvo, P. (2022). Ethically governing artificial intelligence in the field of scientific research and innovation.

[29] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401-4410).

[30] Egan, K. M., Trichopoulos, D., Stampfer, M. J., Willett, W. C., Newcomb, P. A., Trentham-Dietz, A., ... & Baron, J. A. (1996). Jewish religion and risk of breast cancer. *The Lancet*, *347*(9016), 1645-1646.

[31] Brusseau, J. (2022). Using edge cases to disentangle fairness and solidarity in AI ethics. *AI and Ethics*, *2*(3), 441-447.

[32] Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., ... & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*.

[33] Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4-1.

[34] Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., ... & Nascimento, E. G. (2023). Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, *7*(1), 15.

[35] Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a standard for identifying and managing bias in artificial intelligence. *NIST special publication*, *1270*(10.6028).

[36] There's More to AI Bias Than Biased Data, Report Highlights. (2022b, March 16). Retrieved from https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights

[37] Dua, D., & Graff, C. (2017). UCI Machine Learning Repository. University of California, Irvine. *School of Information and Computer Sciences*.

[38] Rhue, L., & Clark, J. (2020). Automatically Signaling Quality? A Study of the Fairness-Economic Tradeoffs in Reducing Bias through AI/ML on Digital Platforms. *A Study of the Fairness-Economic Tradeoffs in Reducing Bias through AI/ML on Digital Platforms (January 10, 2020). NYU Stern School of Business*.

[39] Merler, M., Ratha, N., Feris, R. S., & Smith, J. R. (2019). Diversity in faces. *arXiv preprint arXiv:1901.10436*.

[40] Redmond, M. (2011). Communities and crime unnormalised data set. *UCI Machine Learning Repository. In website: http://www. ics. uci. edu/mlearn/MLRepository. html*.

[41] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

[42] Elizabeth, S. A. (2017). What is the Point of Equality?. In *Theories of Justice* (pp. 133-183). Routledge.

[43] Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

[44] Bias and unfairness in data-informed decisions (2013, August 8). Retrieved from https://www.futurelearn.com/info/courses/data-science-artificial-intelligence/0/steps/147783

[45] Li, J., & Chignell, M. (2022). FMEA-AI: AI fairness impact assessment using failure mode and effects analysis. *AI and Ethics*, 1-14.

[46] Discrimination and biases - Ethics of AI. (n.d.-b). Retrieved from https://ethics-of-ai.mooc.fi/chapter-6/3-discrimination-and-biases/

[47] Alelyani, S. (n.d.-b). Detection and Evaluation of Machine Learning Bias. Retrieved from https://www.mdpi.com/2076-3417/11/14/6271

[48] Machine Learning Glossary: Fairness |. (n.d.-b). Retrieved from https://developers.google.com/machine-learning/glossary/fairness

[49] Ombudsman, F. W. (2020). Protection from discrimination at work.

[50] Dabas, A. (2021, December 15). Bias in Artificial Intelligence - Abhishek Dabas. Retrieved from https://adabhishekdabas.medium.com/bias-in-artificial-intelligence-d2ccec3abb2b

[51] Gan, I., & Moussawi, S. (2022, January). A Value Sensitive Design Perspective on AI Biases. In *HICSS* (pp. 1-10).

[52] Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. In *Early engagement and new technologies: Opening up the laboratory* (pp. 55-95). Springer, Dordrecht.

[53] Victor, D. (2016). Microsoft created a Twitter bot to learn from users. It quickly became a racist Jerk. *The New York Times*.

[54] O'neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

[55] Bughin, J., Seong, J., Manyika, J., Chui, M., & Joshi, R. (2018). Notes from the AI frontier: Modeling the impact of AI on the world economy. *McKinsey Global Institute*, *4*.

[56] Luengo-Oroz, M. (2019). Solidarity should be a core ethical principle of AI. *Nature Machine Intelligence*, *1*(11), 494-494.

[57] Nascimento, A. M., da Cunha, M. A. V. C., de Souza Meirelles, F., Scornavacca Jr, E., & De Melo, V. V. (2018). A Literature Analysis of Research on Artificial Intelligence in Management Information System (MIS). In *AMCIS*.

[58] Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimisation* (pp. 1-9).

[59] Ricardo, B. Y. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54-61.

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

[61] Barton, C., Chettipally, U., Zhou, Y., Jiang, Z., Lynn-Palevsky, A., Le, S., ... & Das, R. (2019). Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. *Computers in biology and medicine*, *109*, 79-84.

[62] Atkinson, K., Bench-Capon, T., & Bollegala, D. (2020). Explanation in AI and law: Past, present and future. *Artificial Intelligence*, *289*, 103387.

[63] Ho, D. A., & Beyan, O. (2020). Biases in data science lifecycle. *arXiv preprint arXiv:2009.09795*.

[64] Fink, A. (2019). *Conducting research literature reviews: From the internet to paper*. Sage publications.

[65] Garg, A., & Sl, DR (2021). PCIV method for Indirect Bias Quantification in AI and ML Models.

[66] Floridi, L. (Ed.). (2010). *The Cambridge handbook of information and computer ethics*. Cambridge University Press.

[67] Simon, J., Wong, P. H., & Rieder, G. (2020). Algorithmic bias and the Value Sensitive Design approach. *Internet Policy Review*, *9*(4), 1-16.

[68] Dexe, J., Franke, U., Nöu, A. A., & Rad, A. (2020, July). Towards increased transparency with value sensitive design. In *International Conference on Human-Computer Interaction* (pp. 3-15). Cham: Springer International Publishing.

[69] Loureiro, T. P. P. R. B., Lisboa, F. V. N., Cruz, G. O. R., Peixoto, R. M., de Sousa Guimarães, G. A., dos Santos, L. L., ... & Nascimento, E. G. S. (2022). BIAS AND UNFAIRNESS IN MACHINE LEARNING MODELS: A SYSTEMATIC LITERATURE REVIEW. *arXiv preprint arXiv:2202.08176*.

[70] Balayn, A., Lofi, C., & Houben, G. J. (2021). Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate

684 bias and unfairness within data management and analytics systems. *The VLDB Journal*, *30*(5),
685 739-768.
686 [71] Huang, J., Galal, G., Etemadi, M., & Vaidyanathan, M. (2022). Evaluation and mitigation of racial
687 bias in clinical machine learning models: scoping review. *JMIR Medical Informatics*, *10*(5),
688 e36388.
689 [72] Orphanou, K., Otterbacher, J., Kleanthous, S., Batsuren, K., Giunchiglia, F., Bogina, V., ... &
690 Kuflik, T. (2021). Mitigating Bias in Algorithmic Systems-A Fish-Eye View. *ACM Computing*
691 *Surveys (CSUR)*.
692 [73] Walch, K. (2020, November 23). 9 data quality issues that can sideline AI projects. Retrieved from
693 https://www.techtarget.com/searchenterpriseai/feature/9-data-quality-issues-that-can-sideline-AI-
694 projects
695 [74] Acharyya, R., Das, S., Chattoraj, A., Sengupta, O., & Tanveer, M. I. (2020). Detection and
696 Mitigation of Bias in Ted Talk Ratings. *arXiv preprint arXiv:2003.00683*.
697 [75] Srinivasu, P. N., Sandhya, N., Jhaveri, R. H., & Raut, R. (2022). From blackbox to explainable AI
698 in healthcare: existing tools and case studies. *Mobile Information Systems*, *2022*, 1-20.
699 [76] Basereh, M., Caputo, A., & Brennan, R. (2021, September). Fair ontologies for transparent and
700 accountable ai: A hospital adverse incidents vocabulary case study. In *2021 Third International*
701 *Conference on Transdisciplinary AI (TransAI)* (pp. 92-97). IEEE.
702 [77] Qiang, V., Rhim, J., & Moon, A. (2023). No such thing as one-size-fits-all in AI ethics frameworks:
703 a comparative case study. *AI & SOCIETY*, 1-20.
704 [78] Chhillar, D., & Aguilera, R. V. (2022). An eye for artificial intelligence: Insights into the governance
705 of artificial intelligence and vision for future research. *Business & Society*, *61*(5), 1197-1241.
706 [79] Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a standard
707 for identifying and managing bias in artificial intelligence. *NIST special publication*, *1270*(10.6028).
708 [80] Madaio, M., Egede, L., Subramonyam, H., Wortman Vaughan, J., & Wallach, H. (2022). Assessing
709 the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for
710 Support. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1-26.
711 [81] Peters, U. (2022). Algorithmic political bias in artificial intelligence systems. *Philosophy &*
712 *Technology*, *35*(2), 25.
713 [82] Landers, R. N., & Behrend, T. S. (2022). Auditing the AI auditors: A framework for evaluating
714 fairness and bias in high stakes AI predictive models. *American Psychologist*.
715 [83] Srinivasan, R., & Chander, A. (2021). Biases in AI systems. *Communications of the ACM*, *64*(8),
716 44-49.
717 [84] Pospielov, S. (2022, June 20). How To Reduce Bias in Machine Learning. Retrieved from
718 https://www.spiceworks.com/tech/artificial-intelligence/guest-article/how-to-reduce-bias-in-
719 machine-learning/

[85] Barba, P. (2021, March 3). 6 Ways to Combat Bias in Machine Learning. Retrieved from https://builtin.com/machine-learning/bias-machine-learning

[86] Schmelzer, R. (2020, June 10). 6 ways to reduce different types of bias in machine learning. Retrieved from https://www.techtarget.com/searchenterpriseai/feature/6-ways-to-reduce-different-types-of-bias-in-machine-learning

[87] Shestakova, V. (2021). Best Practices to Mitigate Bias and Discrimination in Artificial Intelligence. *Performance Improvement*, *60*(6), 6-11.

[88] Stoyanovich, J., Howe, B., & Jagadish, H. V. (2020). Responsible data management. *Proceedings of the VLDB Endowment*, *13*(12).

[89] Seymour, W. (2018). Detecting bias: does an algorithm have to be transparent in order to Be Fair?. *BIAS 2018*.

[90] Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1356.

[91] Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., & Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e1452.

[92] Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, *169*(12), 866-872.

[93] Richardson, B., & Gilbert, J. E. (2021). A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions. *arXiv preprint arXiv:2112.05700*.

[94] Benjamins, R., Barbado, A., & Sierra, D. (2019). Responsible AI by design in practice. *arXiv preprint arXiv:1909.12838*.

[95] Caldwell, S., Sweetser, P., O'Donnell, N., Knight, M. J., Aitchison, M., Gedeon, T., ... & Conroy, D. (2022). An agile new research framework for hybrid human-AI teaming: Trust, transparency, and transferability. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *12*(3), 1-36.

[96] Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & Van Moorsel, A. (2020, January). The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 272-283).

[97] de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government information quarterly*, *39*(2), 101666.

[98] Clarke, R. (2019). Regulatory alternatives for AI. *Computer Law & Security Review*, *35*(4), 398-409.

[99] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature machine intelligence*, *1*(9), 389-399.

[100] Vasudevan, S., & Kenthapadi, K. (2020, October). Lift: A scalable framework for measuring fairness in ml applications. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 2773-2780).

[101] Berente, N., Gu, B., Recker, J., & Santhanam, R. (2019). Managing Ai. *MIS Quarterly*, 1-5.

[102] Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019, January). Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 339-348).

[103] Madaio, M., Egede, L., Subramonyam, H., Wortman Vaughan, J., & Wallach, H. (2022). Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1-26.

[104] Umbrello, S., & Van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics*, *1*(3), 283-296.

[105] Belenguer, L. (2022). AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *AI and Ethics*, 1-17.

[106] Celi, L. A., Cellini, J., Charpignon, M. L., Dee, E. C., Dernoncourt, F., Eber, R., ... & Yao, S. (2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digital Health*, *1*(3), e0000022.

[107] Aera (iPosterSessions - an aMuze! Interactive system). (n.d.). Retrieved from https://aera22-aera.ipostersessions.com/Default.aspx?s=0F-E4-C1-0C-7F-64-F5-D8-26-42-67-7C-A6-EB-B7-DC

[108] Dankwa-Mullan, I., & Weeraratne, D. (2022). Artificial Intelligence and Machine Learning Technologies in Cancer Care: Addressing Disparities, Bias, and Data Diversity. *Cancer Discovery*, *12*(6), 1423-1427.

[109] Delgado, J., de Manuel, A., Parra, I., Moyano, C., Rueda, J., Guersenzvaig, A., ... & Puyol, A. (2022). Bias in algorithms of AI systems developed for COVID-19: A scoping review. *Journal of Bioethical Inquiry*, 1-13.

[110] Curto, G., Jojoa Acosta, M. F., Comim, F., & Garcia-Zapirain, B. (2022). Are AI systems biased against the poor? A machine learning analysis using Word2Vec and GloVe embeddings. *AI & society*, 1-16.

[111] Gupta, M., Parra, C. M., & Dennehy, D. (2021). Questioning racial and gender bias in AI-based recommendations: Do espoused national cultural values matter?. *Information Systems Frontiers*, 1-17.

[112] Pethig, F., & Kroenung, J. (2022). Biased humans,(un) biased algorithms?. *Journal of Business Ethics*, 1-16.

[113] Johansen, J., Pedersen, T., & Johansen, C. (2021). Studying human-to-computer bias transference. *AI & SOCIETY*, 1-25.

[114]Feuerriegel, S., Dolata, M., & Schwabe, G. (2020). Fair AI: Challenges and Opportunities. Business & Information Systems Engineering, 62 (4), 379–384.

[115]John-Mathews, J. M., Cardon, D., & Balagué, C. (2022). From reality to world. A critical perspective on AI fairness. *Journal of Business Ethics*, 1-15.

[116]Turney, P. (1995). Bias and the quantification of stability. *Machine Learning*, *20*(1), 23-33.

[117]Michael, K., Abbas, R., Jayashree, P., Bandara, R. J., & Aloudat, A. (2022). Biometrics and AI Bias. *IEEE Transactions on Technology and Society*, *3*(1), 2-8.

[118]Zhang, Z., Li, J., Stork, D. G., Mansfield, E., Russell, J., Adams, C., & Wang, J. Z. (2022). Reducing Bias in AI-based Analysis of Visual Artworks. *IEEE BITS the Information Theory Magazine*.

[119]Suri, J. S., Agarwal, S., Jena, B., Saxena, S., El-Baz, A., Agarwal, V., ... & Naidu, S. (2022). Five Strategies for Bias Estimation in Artificial Intelligence-based Hybrid Deep Learning for Acute Respiratory Distress Syndrome COVID-19 Lung Infected Patients using AP (ai) Bias 2.0: A Systematic Review. *IEEE Transactions on Instrumentation and Measurement*.

[120]Luengo-Oroz, M., Bullock, J., Pham, K. H., Lam, C. S. N., & Luccioni, A. (2021). From artificial intelligence bias to inequality in the time of COVID-19. *IEEE Technology and Society Magazine*, *40*(1), 71-79.

[121]Straw, I. (2020). The automation of bias in medical Artificial Intelligence (AI): Decoding the past to create a better future. *Artificial intelligence in medicine*, *110*, 101965.

[122]Hildebrandt, M. (2021). The issue of bias. The framing powers of machine learning.

[123]Society to Improve Diagnosis in Medicine. (2021, May 24). Bias in Artificial Intelligence. Retrieved from https://www.improvediagnosis.org/publications/improvedx-may-2021/bias-in-artificial-intelligence/

[124]Yavuz, C. (2019). Machine Bias: Artificial Intelligence and Discrimination.

[125]Zuiderveen Borgesius, F. (2018). Discrimination, artificial intelligence, and algorithmic decision-making.

[126]Cakir, C. (2020). Fairness, Accountability and Transparency–Trust in AI and Machine Learning. *The LegalTech Book: The Legal Technology Handbook for Investors, Entrepreneurs and FinTech Visionaries*, 35-37.

[127]Agarwal, A., Agarwal, H., & Agarwal, N. (2022). Fairness Score and process standardisation: framework for fairness certification in artificial intelligence systems. *AI and Ethics*, 1-13.

Table 1(on next page)

Few popular datasets, along with their characteristics

1

Dataset Name	Attributes	No.of Records	Area
COMPAS[33]	Criminal Histories, Jail & Prison Times, Demographics ,COMPAS Risk Scores	18,610	Social
German Credit[34]	Housing Status , Personal Status, Amount , Credit Score, Credit, Sex	1,000	Financial
UCI Adult[35]	Age, Race, Hours-Per-Week , Marital Status, Occupation,Education, Sex, Native Country	58,842	Social
Diversity in Faces[36]	Age, Pose, Facial Symmetry And Contrast, Craniofacial Distances , Gender, Skin Color, Resolution Along With Diverse Areas And Ratios	1 million	Facial images
Communities and Crime[37]	Crime & Socio-Economic Data	1,994	Social
Winobias[38]	Male Or Female Stereotypical Occupations	3,160	Coreference resolution
Recidivism in Juvenile Justice [39]	Juvenile Offenders' Data And Prison Sentences	4,753	Social
Pilot Parliaments Benchmark[40]	National Parliaments Data (e.g., Gender and Race)	1,270	Facial images

2

3

Table 2(on next page)

A socio-technical approach for bias mitigation across the CRISP-DM development cycle

1

	Business Understanding	Data Understanding	Data Preparation	Modelling	Evaluation	Deployment
Measurement Bias	Team diversity, Exchange with domain expert	Proxy estimation	Rapid prototyping			
Social Bias			Learning fair representation, Rapid prototyping, Reweighting, Optimized preprocessing, Data massaging, Disparate impact remover	Adversarial debiasing, Multiple models, Latent variable model, Model interpretability Equalized odds, Prejudice remover		
Sampling Bias				Resampling		Randomness
Representation Bias	Team diversity	Data plotting, Exchange with domain experts	Reweighing, Data augmentation	Model interpretability		
Negative Bias			Cross dataset generalization	Bag of words		
Label Bias		Exchange with domain experts	Data massaging			
Sample Selection Bias			Reweighing			
Confounding Bias						Randomness
Design Bias			Rapid prototyping	Exchange with domain experts, Resampling, Model interpretability, Multitask learning		
Sample Treatment Bias				Resampling	Data augmentation	
Human Evaluation Bias				Resampling	Representative benchmark subgroup validity, Data augmentation	
Test Dataset Bias			Data augmentation			
Deployment Bias	Team diversity, Consequences in context		Rapid prototyping			Monitoring plan, Human supervision
Feedback Bias						Human supervision, Randomness

2

Figure 1

Search words across three predefined domains- AI, ML & SE

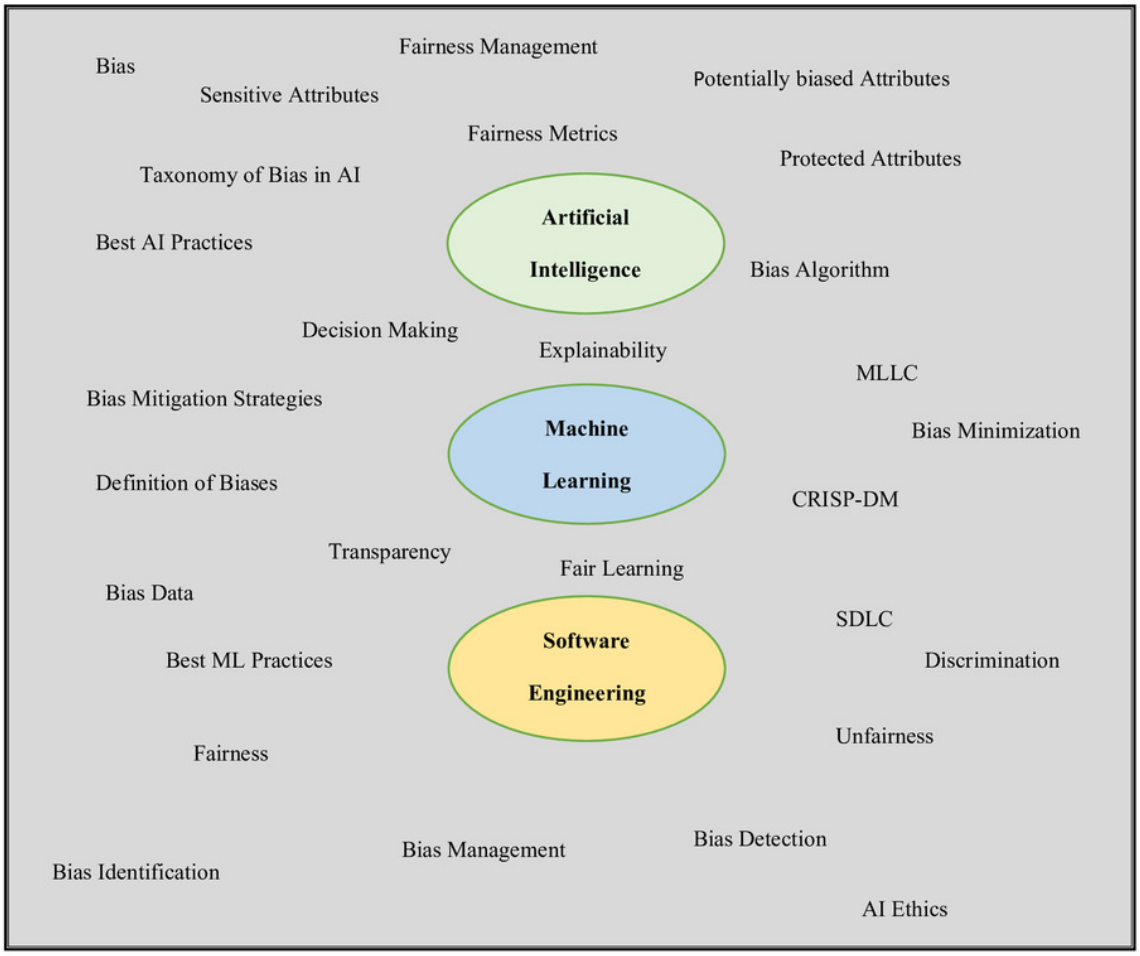
Fairness Management
Bias
Potentially biased Attributes
Sensitive Attributes
Fairness Metrics
Protected Attributes
Taxonomy of Bias in AI
Artificial Intelligence
Best AI Practices
Bias Algorithm
Decision Making
Explainability
MLLC
Bias Mitigation Strategies
Machine Learning
Bias Minimization
Definition of Biases
CRISP-DM
Transparency
Fair Learning
Bias Data
SDLC
Software Engineering
Best ML Practices
Discrimination
Unfairness
Fairness
Bias Management
Bias Detection
Bias Identification
AI Ethics

Figure 2

Trio bias feedback loop among Data, Algorithm and User

Figure 2 Trio bias feedback loop among Data, Algorithm and User.

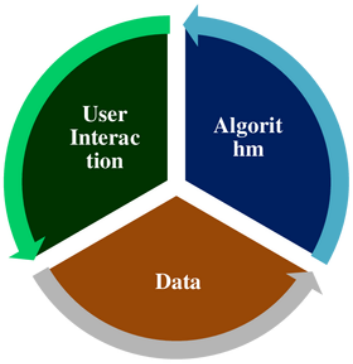

Figure 3

Mapping the SDLC, MLLC and CRISP-DM across different biases categories, minimizing strategies and fairness management

Figure 3 Mapping the SDLC, MLLC and CRISP-DM across different biases categories, minimizing strategies and fairness management.

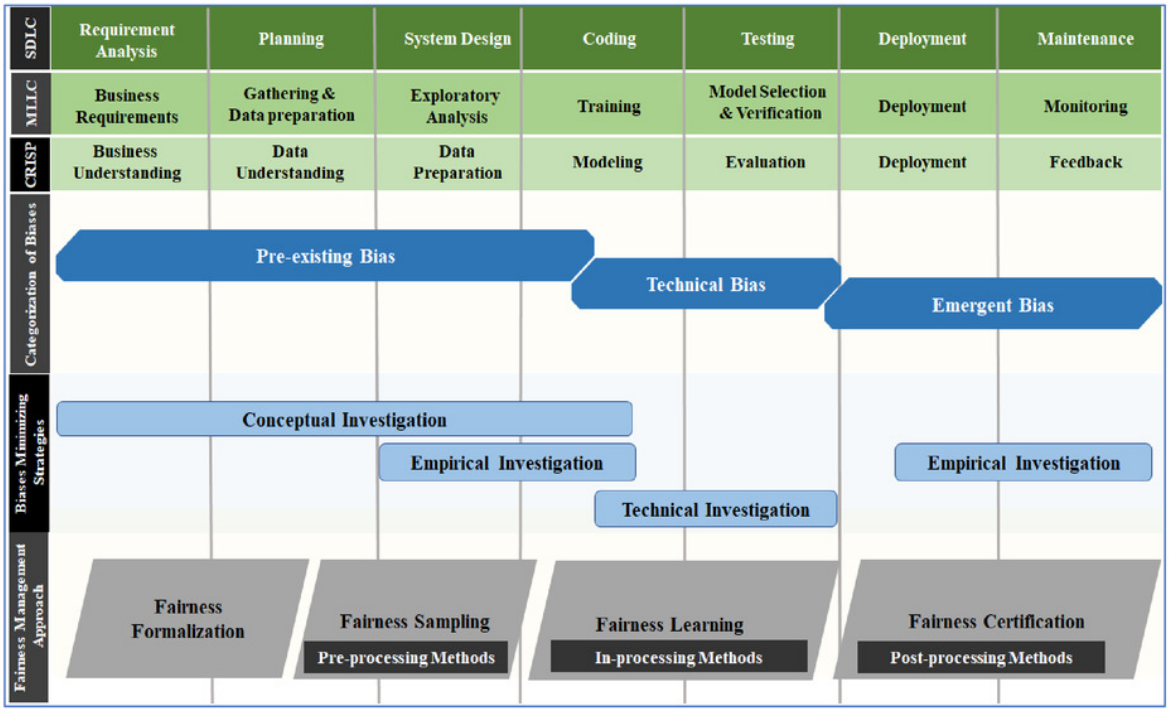

Figure 4

Distribution of biases w.r.t categories across phases of the CRISP-DM development cycle

Figure 4 Distribution of biases w.r.t categories across phases of the CRISP-DM development cycle.

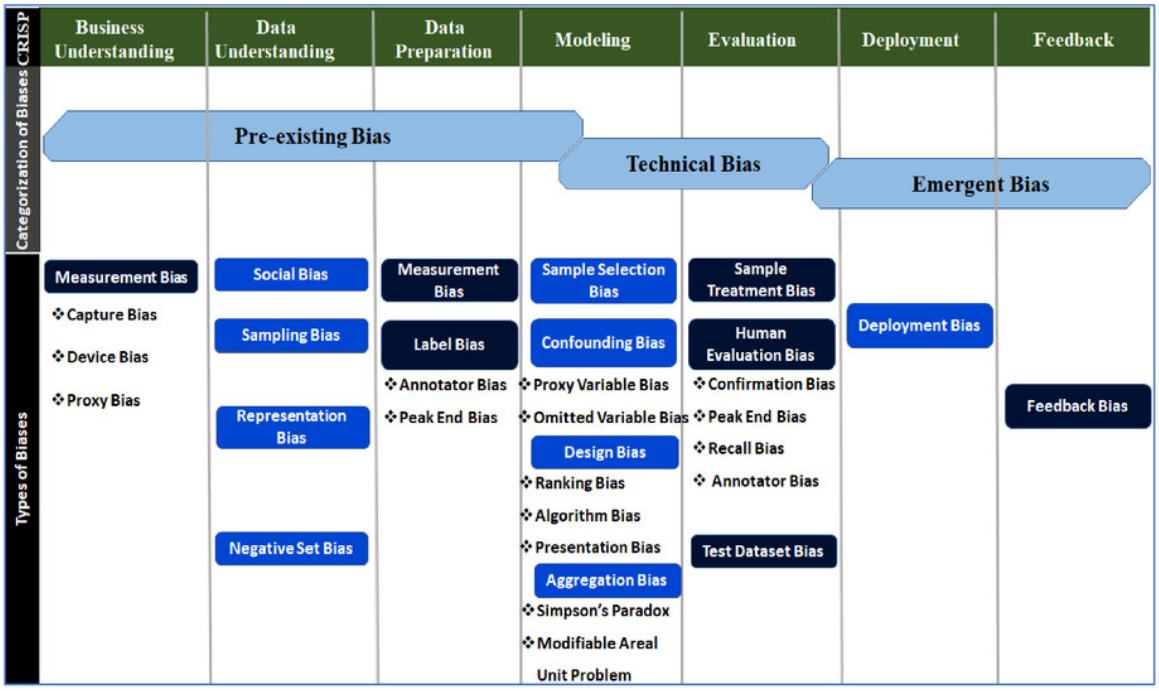

Figure 5

Fairness management approach across multiple phases of the CRISP-DM development cycle
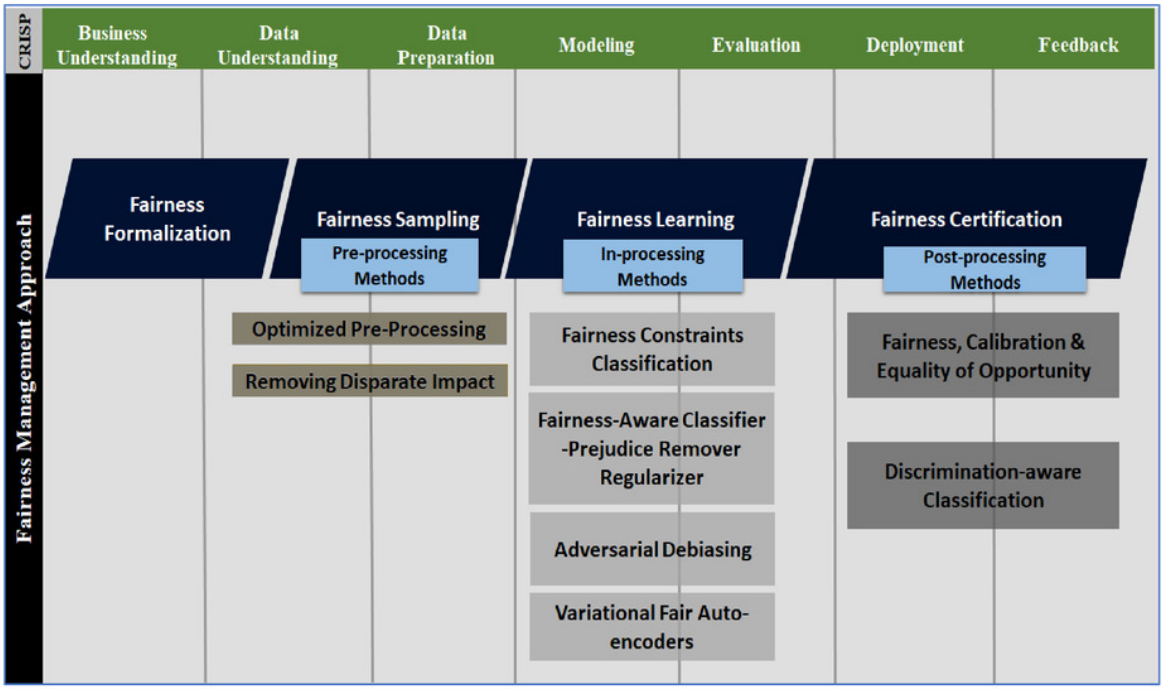
Figure 5 Fairness management approach across multiple phases of the CRISP-DM development cycle.