The Multifaced Nature of Bias in AI: Impact on Model Generalization, Robustness, and Fairness

Eirini Ntoutsi

Abstract

Bias in Artificial Intelligence (AI) is a critical issue that has gained significant attention due to its association with discrimination and harm. Although bias has increasingly carried a negative connotation in recent years, it is not inherently positive or negative. In AI, bias can guide models toward desired outcomes and improve generalization, but it can also lead to discrimination against individuals or groups based on protected characteristics such as gender, race, or age, and undermine model robustness in varying contexts. This chapter explores the multifaceted nature of bias in AI, highlighting its benefits and drawbacks. We discuss how bias can be harnessed to improve models while addressing its negative effects, such as perpetuating inequalities and reducing robustness. The need to understand and manage bias is emphasized to ensure AI systems remain fair, ethical, and effective.

1 Introduction

AI-based systems are widely used to make decisions that impact individuals and society, from screening job applicants to aiding healthcare diagnoses and assessing risks in bail or sentencing, raising concerns about potential bias and human rights issues (Ntoutsi et al., 2020). The discriminative impact of AI-based decision-making on individuals and demographic groups characterized by protected attributes such as gender and race has already

E. Ntoutsi (✉)
Universität der Bundeswehr München, Neubiberg, Deutschland
E-Mail: eirini.ntoutsi@unibw.de

B. Schäffer, F. R. Lieder (Hrsg.), *Maschinen wie wir?*,
https://doi.org/10.1007/978-3-658-48522-1_11

been observed in various cases (West et al., 2019). For instance, the COMPAS system for predicting the risk of re-offending was found to assign higher risk scores to Black defendants and lower scores to White defendants than their actual risk levels- an example of racial bias (Angwin et al., 2022). Similarly, Amazon's resume scanning tool (Dastin, 2018) for ranking job candidates exhibited harmful gender bias by ranking candidates lower if their resumes included gendered terms such as "women's" or references to women's colleges. Such cases naturally raise concerns about the fairness of AI systems. As a result, the term "bias" is often synonymous with discrimination and carries a negative connotation in contemporary AI literature.

Historically, bias was described as "a leaning of the mind" representing an inclination away from a state of indifference.[1] Over time, the term has acquired a more negative connotation, now defined as "an inclination of temperament or outlook, especially a personal and sometimes unreasoned judgement: prejudice".[2] Bias, however, is neither inherently good nor bad, and this applies to both humans and machines. Some human biases can be helpful, such as favoring healthy eating, or some biases can make us more efficient, like starting work early if you're a morning person. On the other hand, biases related to prejudices – favoring or discriminating against a person or group over another in an unfair way – are harmful and should be prohibited, such as declining a job based on gender or race. With respect to human bias, its many facets have been studied by many disciplines (Haselton et al., 2015) including psychology, ethnography, law, and so forth.

Similarly, bias in machines/AI has many facets and is neither good or bad. While most recent discussions on bias focus on discrimination and harmful biases toward individuals or groups, the concept of bias is an old concept in Machine Learning (ML), traditionally referring to the assumptions made by a specific model – the so-called, inductive bias – (Mitchell, 1997), which are necessary to enable generalization from specific instances to broader theories. Other biases guide models in a desired direction in the hypothesis space, for instance bias towards simplicity, such as Occam's Razor, where simpler models are preferred over more complex ones (Domingos, 1999). However, there are also biases like selection bias, that can harm model robustness in new contexts, negatively impacting machine performance, even if no human instances are directly involved – such as training a model on data from a certain context failing to generalize across different contexts.

Rather than demonizing bias, it is crucial to acknowledge its role as a tool for steering AI models in desired directions, while also mitigating harmful biases that lead to discrimination or degrade model performance. The goal of this work is to examine the multifaceted nature of bias, offering a comprehensive overview of both its beneficial and detrimental applications in AI. While a completely bias-free world may be an unrealistic goal for both humans and machines, mastering bias is necessary, especially for the responsible use of AI technology.

[1] https://www.psychologytoday.com/intl/blog/hovercraft-full-eels/202009/biases-are-neither-all-good-nor-all-bad

[2] https://www.merriam-webster.com/dictionary/bias

The rest of the paper is organized as follows: First, we provide a basic introduction of how machines learn. Second, we explore biases that enable ML induction. Next, we examine how biases can undermine model robustness. Afterwards, we delve into harmful biases that may result in discrimination and harm. Finally, we conclude with a summary of the key insights.

2 How Machines Learn

In modern AI, particularly in Machine Learning (ML), computers learn from data without being explicitly programmed, as first articulated by (Samuel, 1959). Unlike traditional programming, where the solution is explicitly coded (e.g., the insertion sort algorithm; (Mehlhorn, 2013), ML derives solutions from patterns within the data, even when the exact solution is unknown or difficult to describe explicitly (e.g., recognizing a cat or detecting early-stage cancer in medical images). These tasks are hard to codify manually due to the complexity and variability of the patterns involved, but ML models can learn intricate patterns from vast amounts of data. Thus, *data* and *learning algorithms* are the essential components of modern AI, enabling machines to implicitly learn and generalize from examples.

Learning from data encompasses different tasks, which vary depending on the level of supervision provided by a (human) expert during the learning process. The primary tasks are: *supervised learning*, where both data and their corresponding labels are available to the learner, with typical examples including classification and regression; *unsupervised learning*, where only the data are available without labels, with typical examples including clustering and dimensionality reduction; and *reinforcement learning*, where supervision is provided in the form of rewards rather than explicit labels for individual instances. In this work, we focus on supervised learning.

Supervised learning involves learning a mapping from input data to output labels based on a set of labelled instances (training data). The input data, typically represented as feature vectors, are paired with their corresponding output labels. The goal is to learn a function or hypothesis or model that can accurately predict the output for new, unseen data. This function is selected from a *hypothesis space*, which represents the set of all possible models that can be learned. The choice of hypothesis space determines the type of model (e.g., linear models, decision trees, neural networks, support vector machines) and the form of the input-output mapping.

The key challenge in supervised learning is to ensure the learned model generalizes well to unseen data. A model that fits the training data too closely might capture noise or specific patterns in the training set, leading to *overfitting*, which results in poor performance over new instances. On the other hand, a model with a too simple hypothesis may fail to capture the underlying structure in the data, resulting in *underfitting*, where the model is unable to capture the complexity of the data. Finding the right model complexity is necessary to prevent overfitting and underfitting, and to improve generalization.

While machines learn from data through algorithms, the choice of data, preprocessing methods, and the selection of algorithms or models are largely driven by human decisions, significantly influencing what and how machines learn, shaping their behavior and outcomes. This applies not only to modern AI, such as ML, but also to traditional AI (Russell & Norvig, 2016). For instance, expert systems rely on human experts to create and update their knowledge base, define "if-then" rules, and interpret the system's results. Similarly, informed search algorithms like A* are driven by heuristics designed by humans based on domain expertise, which estimate how good a particular state is for achieving a goal and guide the search process accordingly.

3 Biases that Enable Machine Learning Induction

Machine Learning often involves inductive learning; *induction* refers to the process of inferring a general model of the domain from a finite set of observations (training data). This also includes the search for this model/hypothesis in a large hypothesis space. *Inductive bias* refers to the set of explicit or implicit assumptions made by a learning algorithm in order to perform induction (Hüllermeier et al., 2013). A classic example of inductive bias is *Occam's Razor*, which expresses a preference for simplicity: given two equally effective models, the simpler one should be preferred for generalization (Domingos, 1999).

Bias-free learning is futile (Mitchell, 1980), meaning that without such bias, induction would not be possible, as the same observations can often be generalized in various ways. As argued in this work, the ability of learning algorithm to generalize relies on incorporating biases that extend beyond strict consistency with the training data, including prior domain knowledge, preferences for simplicity, and consideration of the algorithm's real-world applications. (Montañez et al., 2019) provide a mathematical justification for the necessity of bias in improving learning performance, concluding that biases are essential for better than chance performance. Furthermore, as emphasized in this work, these biases must be correct as their effectiveness depends on how well they align with the actual target being sought.

The concept of inductive bias has been central to AI since the early developments of the field. Early systems, like expert systems, rely on explicit, expert-defined, hard-coded rules (Russell & Norvig, 2016), which often encode human knowledge and its inherent biases. Similarly, informed search algorithms like A* utilize human-designed heuristics to guide their search, reflecting assumptions and preferences introduced by their designers.

Machine Learning models also incorporate biases to balance model complexity and generalization. Different ML algorithms come with distinct inductive biases. Below, we describe some popular learners and their biases; this list of biases is not exhaustive but aims to provide an overview of some of the underlying assumptions made by each learner:

- **Decision Trees (DTs):** A decision tree learner typically selects the first acceptable tree it encounters during its simple-to-complex, hill-climbing search through the space of possible trees. Additionally, it favors shorter trees over longer ones, in line with Oc-

cam's Razor, which improves generalizability. Decision trees also assume that the data space is partitioned into axis-parallel-hyper-rectangles (Mitchell, 1997).

- **Naive Bayes (NBs):** Naive Bayes classifiers, one of the most popular methods for probabilistic induction, assume that all features are conditionally independent given the class label, known as the class-conditional independence assumption, enabling the decomposition of high-dimensional multivariate probabilities into a product. They also assume that all features contribute equally to predictions, ignoring potential correlations between them. The probabilities are based solely on the training data, without any search through the weight space (Langley & Sage, 1994).
- **k-Nearest Neighbors (k-NNs):** k-NNs assume that points close to each other in the feature space have similar outputs (locality assumption) and that all features contribute equally to the distance calculation for computing the neighborhood.
- **Neural Networks (NNs):** Neural network algorithms assume that data have hierarchical structures, where simpler patterns combine to form more complex ones, and that relationships between inputs and outputs can be learned through multiple layers of abstraction or granularity. This reflects a compositional inductive bias (Goyal & Bengio, 2022), where the model assumes that complex tasks can be decomposed into simpler components, allowing it to generalize more effectively by reusing learned features across different contexts.
- **Convolutional Neural Networks (CNNs):** Convolutional neural networks, a specialized type of neural network, make additional assumptions, such as translational invariance (Wang & Wu, 2024), meaning they assume patterns can appear anywhere in an image. This assumption is implemented through convolutional layers and pooling, allowing the model to generalize better by recognizing features regardless of their position.
- **Transformers:** Transformers (Vaswani et al., 2017) assume that dependencies between inputs can be captured regardless of their position, the so-called permutation invariance. This assumption is implemented through self-attention mechanisms, allowing the model to capture long-range dependencies and contextual relationships without relying on locality.

Towards Generalizable Modes: Bias-Variance Trade-Off

The bias-variance trade-off involves balancing bias, the model's assumptions that may cause underfitting, and variance, how small data changes affect performance, leading to overfitting. High bias models make strong assumptions and may underfit the training data, failing to capture the underlying patterns. High variance models with little bias, may fit the training data too closely, capturing noise and resulting in poor generalization to new data. Examples of low bias, high variance models include fully grown decision trees, which can perfectly fit the training data but may overfit; k-NNs with small k, where the model memorizes the data instances and is highly sensitive to noise and high-degree polynomial regression, which can fit every point in the training set but performs poorly on new data.

Techniques such as regularization (Tian & Zhang, 2022) help manage the bias-variance trade-off by controlling model complexity. Regularization methods like L2 regularization,

commonly used in neural networks (Goodfellow et al., 2016), penalize large weights, reducing variance and improving generalization. Ensemble methods such as bagging and boosting also reduce variance in high-bias models by averaging the outputs of multiple weak learners, leading to a more balanced performance (Gupta et al., 2022). Understanding and managing the inductive biases of each model is crucial for selecting the right model for a given problem and designing models aligned with specific challenges, thereby enabling effective machine learning.

4 Biases that Undermine Model Robustness

In the previous section, we discussed how inductive biases enable generalization by introducing assumptions that guide learning. However, biases are not limited to learning algorithms. Data-related biases, arising from the way data are collected, labeled, preprocessed etc., can significantly undermine model robustness, that is, model's ability to perform well across diverse contexts. These biases often lead to poor generalization. As the use of AI becomes more widespread, addressing these data-related biases becomes crucial.

Several works describe biases in specific application domains; for example, (Fabbrizzi et al., 2022) provide a detailed discussion of such biases in computer vision. Below, we describe key biases that can negatively impact a model's robustness, with examples from the agriculture domain, specifically crop classification using satellite images.[3]

- **Sampling or Selection bias:** Sampling bias occurs when the training data is not representative of the entire population or the real-world distribution where the model will be deployed. This can lead to models that perform well on the specific training set but fail to generalize to new, unseen data.
- In crop prediction, sampling bias may occur if the training data are predominantly sourced from a specific region or climate, leading to poor performance in other regions with different climate, soil, etc. characteristics. For example, a crop prediction model trained on data from Germany may struggle when deployed in Greece, where crops look different due to the warmer climate.
- **Class imbalances:** Class-imbalance bias is a specific form of sampling bias, commonly seen in ML. It occurs when certain classes are over-represented in the training data, leading to skewed predictions that favor majority class(es) and may overlook the minority class(es).
- In crop prediction, the distribution of crop types is typically heavily skewed, with many types falling into the long tail. If not properly addressed, the prediction rate for these long tail classes will be significantly lower.

[3] The crop prediction application is part of a use case within the EU project STELAR. (grant agreement No. 101070122), aimed at creating a knowledge data lake management system for agriculture, harnessing AI and big data technologies.

- **Label bias:** Labelled data are essential for training models in supervised learning tasks, and the quality of these labels is crucial. Label bias, as defined by (Jiang & Nachum, 2020), occurs when the labels systematically differ from the ground truth, often due to labelling errors or inconsistencies.
- In the crop prediction case, the satellite images are not labelled; labels are typically automatically assigned based on the EuroCrops dataset, which contains geo-referenced polygons of agricultural croplands from 16 countries in the European Union (EU), along with information on the specific crop species grown in those regions (Schneider et al., 2023).
- **Measurement bias:** Measurement bias occurs when data collected has inaccuracies or it is systematically flawed due to issues with the measurement process, leading to unreliable input features and affecting model performance.
- In agriculture, measurement bias manifests when Leaf Area Index (LAI) values from satellite images are impacted by cloud coverage, leading to invalid measurements, particularly during cloud-heavy periods.
- **Domain-specific practices bias:** Domain-specific biases arise when practices or assumptions unique to a particular field lead to systematic errors in data collection or labelling, affecting model performance.
- In the agriculture domain, the labelling process via EuroCrops (Schneider et al., 2023) mentioned above, assigns fixed crop labels annually. This can lead to label bias over time, as it may overlook important factors like crop rotation practices, sudden disease or pest outbreaks, and other dynamic agricultural changes.

Towards Robust Models

Biases like those described above can hinder model generalization in real-world scenarios. Identifying and addressing these biases is essential for building robust models. Techniques such as imbalance learning to tackle class imbalances (Japkowicz & Stephen, 2002) can mitigate these issues. Additionally, dataset documentation, like datasheets for datasets (Gebru et al., 2018), provides transparency about data provenance, collection, and biases, helping practitioners make informed decisions and mitigate risks. However, even when it is not possible to completely eliminate such biases, being aware of their presence and interpreting model results accordingly remains critical.

5 How Machines Can Be Unfair: Biases that Lead to Discrimination and How to Mitigate Them

In the previous sections, we explored how inductive biases enable learning and generalization and how data-related biases can undermine model robustness. Beyond impacting generalization and robustness, biases in AI systems can also lead to unfair outcomes and perpetuate discrimination. These biases often arise from societal inequalities embedded in data, design choices within AI pipelines, or unintended consequences of optimization ob-

jectives. This section first examines how AI systems can result in unfair outcomes and discrimination, and then outlines fairness-aware learning approaches to mitigate these biases.

5.1 Why Machines Discriminate: The Data and Algorithm Bias Behind AI Discrimination

Since AI systems rely heavily on data, the primary source of bias in AI comes from *the data* themselves. Data is typically generated by humans (e.g., social media posts) or collected through systems created by humans. As a result, whatever biases exist in humans can be embedded into AI systems, as has been demonstrated in various cases, such as gender bias in word embeddings (Bolukbasi et al., 2016), racial bias in computer vision algorithms (Fabbrizzi et al., 2022) and bias in large language models (Gallegos et al., 2024). Even worse, human biases can be amplified due to the complex sociotechnical nature of these systems, such as the Web (Berendt et al., 2021). Additionally, AI systems typically rely on complex pipelines and feedback loops that can further amplify and propagate bias, or even create new forms of bias such as text length bias in machine translation (Murray & Chiang, 2018). While this work does not aim to fully cover bias sources, interested readers are referred to comprehensive surveys on bias in AI (Ntoutsi et al., 2020), including data type-specific surveys such as those on NLP (Blodgett et al., 2020), computer vision (Fabbrizzi et al., 2022), multimodal data (Adewumi et al., 2024) and application-specific surveys in areas like hiring (Fabris et al., 2024) and education (Baker & Hawn, 2022).

The second key component of AI systems is the *learning algorithm*, which is designed to optimize specific objectives related to the learning task (Mitchell, 1997). For instance, in binary supervised learning, the goal is typically to improve the separation between positive and negative classes, often modelled as minimizing empirical risk using a loss function that measures the error between predicted and actual labels. The optimization problem aims to find the best model (hypothesis) that minimizes this error on the training data within the hypothesis space of possible models. However, traditional algorithms do not incorporate fairness as part of their objectives and therefore overlook the performance disparities across different demographic groups.

As a result, the *AI models* learned reflect the complex interactions between data and algorithms, which can lead to unintended "shortcuts" or biases. For example, a hiring algorithm might make decisions based on gender, even if gender is not explicitly used, by relying on proxy attributes (Prince & Schwarcz, 2019). Similarly, a sentiment analysis algorithm might make decisions based on group identifiers, such as race, while ignoring the actual context of the text (Kennedy et al., 2020).

Theses biases can lead to actual harm, which can be categorized into allocative and representational harm (Whittaker et al., 2018). Allocative harms refer to the unfair withholding of access to services, resources, or opportunities for individuals or groups, such as predictive policing disproportionately targeting minority communities (Jones, 2020). Representational harms occur when AI systems reproduce and amplify harmful stereotypes like language models reinforcing gender stereotypes in professions (Bolukbasi et al., 2016).

5.2 Mitigating Bias and Discrimination: Fairness-Aware Machine Learning

To address bias in AI systems, the field of fairness-aware learning has recently emerged, aiming to ensure responsible AI development. This young, interdisciplinary domain focuses on creating AI that does not discriminate on the basis of protected attributes such as gender and race. While bias and fairness have long been studied in fields like philosophy (Kelly, 2022) and law (Fredman, 2022), fairness-aware learning in AI is relatively new, with the seminal work of Pedreschi et al. (2008). Despite its novelty, a substantial body of work has already been proposed, broadly categorized into three areas (Ntoutsi et al., 2020): understanding bias, mitigating bias and accounting for bias, briefly described after.

5.2.1 Understanding Bias and (Un)fairness

Research in this category explores how bias is created in society and enters sociotechnical systems (Berendt et al., 2021). As noted in Section 3.1, the two main sources of bias in AI are data and algorithms. Works in this category investigate finer mechanisms of bias creation and propagation (Srinivasan & Chander, 2021). Significant research focuses on how bias manifests in data through protected attributes (Yu et al., 2021), proxies (Yeom et al., 2018), or the under-/over-representation of demographic groups (Roy et al., 2022). Studies also explore bias manifestation for specific data types like images (Fabbrizzi et al., 2022), text (Gallegos et al., 2024), and multimodal data (Adewumi et al., 2024). Additionally, substantial work addresses the definition and measurement of unfairness (Verma & Rubin, 2018). Many fairness definitions originate from social sciences or law and are operationalized in AI (Hutchinson & Mitchell, 2019). However, fairness cannot be reduced to a single definition; it is context dependent (e.g., spatial, temporal, or application-specific) with different value systems requiring different mechanisms for fair decision making. Moreover, theoretical results demonstrate that multiple definitions of fairness are often mutually exclusive, making it impossible to satisfy all fairness criteria simultaneously (Friedler et al., 2021).

Fairness measures are broadly categorized into individual and group fairness measures (Verma & Rubin, 2018). Group fairness measures compare model performance between protected and non-protected groups. In this category belong measures such as demographic parity (Dwork et al., 2012), equalized odds (Hardt et al., 2016) and predictive rate parity (Zafar et al., 2017). In contrast, individual fairness measures focus on whether similar individuals receive similar treatments from the model. Measures such as counterfactual fairness (Kusner et al., 2017) belong to this category. Most fairness measures have been designed for binary classification tasks with a single binary protected attribute, although research also extends these measures to multi-class problems, multiple protected attributes (Roy et al., 2023), and other learning tasks such as clustering (Chhabra et al., 2021).

5.2.2 Bias Mitigation and Debiasing Techniques

Given the importance of the topic, bias mitigation in AI systems has attracted significant attention. Various approaches to bias mitigation have been proposed: some focus on the

data (pre-processing approaches), some on the learning algorithms (in-processing approaches), and others on the learned ML models (post-processing approaches). Moreover, hybrid approaches, including end-to-end methods, have been proposed to address bias more holistically at multiple stages of the AI pipeline.

Pre-processing approaches: The intuition behind many preprocessing methods is that making the data "more fair" will result in a "less unfair" model. For many methods focusing on group fairness, the key idea is to balance the protected and non-protected groups in the dataset, while adhering to the design principle of minimal data interventions to retain the utility of the data for the learning task. Several techniques have been proposed for this purpose, including: instance selection methods, where specific data points are sampled to achieve balance (Kamiran et al., 2010); instance weighting methods, which adjust the influence of different groups by assigning weights to instances (Calders et al., 2009); class modification methods known as "massaging" methods, where class labels of certain instances are altered to reduce bias (Kamiran & Calders, 2009; Luong et al., 2011), and data augmentation methods that generate synthetic instances to improve dataset balance (Iosifidis & Ntoutsi, 2018).

Most of the pre-processing techniques are heuristic, meaning their impacts are not always well controlled. More principled approaches, such as those proposed by (Calmon et al., 2017) aim to offer a more structured and theoretically grounded framework for bias mitigation. A key advantage of these methods is their high versatility, as they are algorithm- and model-agnostic, allowing "de biased" datasets to be used with any learning algorithm.

In-processing approaches: The intuition behind in-processing methods is that working directly with the learning algorithm allows for better control over the model's behavior. The key idea is to explicitly incorporate the model's discriminatory behavior into the objective function, aiming to achieve both predictive performance and fairness. Several techniques have been developed, including: regularization methods that penalize discrimination within the learning process (Kamishima et al., 2012; Zhang & Ntoutsi, 2019; Dwork et al., 2012; Padala & Gujar, 2020); adding fairness constraints that directly restrict the model's discriminatory effects (Zafar et al., 2017); training on fair latent target labels (Krasanakis et al., 2018); adversarial training where an adversary attempts to predict the protected attribute (Zhang et al., 2018); compositional methods that decouple the groups and train separate models for each group (Dwork et al., 2018); and in-training distribution alteration in ensemble models (Iosifidis & Ntoutsi, 2019).

Several studies explicitly model a trade-off between fairness and accuracy in machine learning models, for instance (Padala & Gujar, 2020), as reducing discrimination can sometimes conflict with the goal of maximizing predictive accuracy. Other works argue that there is no trade-off (Dutta et al., 2020) and enforcing fairness might improve model accuracy on unbiased test data (Maity et al., 2020). There are also approaches that learn to select the objective (accuracy or fairness) to optimize at each step of the optimization process (Roy & Ntoutsi, 2022). In-processing approaches are popular for bias mitigation, but are model- and algorithm-specific, requiring the development of new methods or adaptations for different algorithms.

Post-processing approaches: Post-processing approaches are applied post-hoc, after the model has been optimized for predictive performance. The key idea is to first train the model for predictive performance, as in standard ML, and then apply adjustments to improve fairness. These adjustments should involve minimal interventions with the aim of preserving predictive performance while improving fairness. There are two main types of post-processing methods for bias mitigation: white-box approaches, which alter the model's internal structure, and black-box approaches, which adjust the model's predictions without modifying the internal workings (such as neural networks and other black-box models). White box approaches work with interpretable models, examples include correcting the confidence of classification rules (Pedreschi et al., 2019), probabilities in Naive Bayes models (Calders & Verwer, 2010) or the class label of leaves in decision trees (Kamiran et al., 2010). Black-box approaches work with black-box models and aim at post-hoc adjustment of the decision boundary. Examples include wrapping a fair classifier on top of a black-box base classifier (Agarwal et al., 2018), promoting and demoting predictions close to the decision boundary (Kamiran et al., 2018), and differentiating the decision boundary over groups (Hardt et al., 2016).

Post-processing methods prioritize predictive performance, with fairness considered as a secondary objective. Moreover, these methods are model-specific, meaning that new models require the development of new methods or adaptations. However, black-box approaches can still be useful in practice because we often have access only to the model's outcomes, not its training process.

Hybrid approaches: Pre-processing approaches focus exclusively on the data, in-processing approaches on the algorithm, and post-processing approaches on the model. In contrast, hybrid approaches address discrimination more holistically by integrating interventions across data, algorithms, and models. For instance, Hu et al. (2020) addresses bias and discrimination in feature representation and classification tasks using an autoencoder to obfuscate protected attribute information and a fairness-aware classifier. While most bias-mitigation approaches focus on a single protected attribute, methods also exist for addressing multi-attribute discrimination, such as those in Roy et al., 2022. The challenge increases as the optimization problem becomes more complex with multiple objectives, and multi-attribute scenarios often involve group imbalances and extreme class imbalances within those groups (Roy et al., 2023; Brzezinski et al., 2024).

5.2.3 Accounting for Bias

Methods in this category address bias either proactively or retroactively. Proactive approaches include bias-aware data collection (Fabbrizzi et al., 2022) like the Pilot Parliaments Benchmark (PPB) (Buolamwini & Gebru, 2018) dataset, which consists of photos of members from six national parliaments, collected to ensure a balanced representation of gender and skin color among the subjects. Other approaches involve documenting and describing bias, drawing inspiration from methods used to document data, such as datasheets for datasets (Gebru et al., 2021), and machine learning models, like model cards (Mitchell et al., 2019). Various formalisms, including ontologies (Russo & Vidal, 2024), are employed for this purpose. Such a documentation is crucial as it promotes transparency and facilitates accountability.

Retroactive approaches include explanation methods aimed at clarifying algorithmic decisions and revealing potential biases in decision-making. These explanations are essential for assessing whether decisions are biased (Fragkathoulas et al., 2024; Deck et al., 2023) and can be even used to debias (Cai et al., 2022; Kim et al., 2024) or correct the system (Weber et al., 2023). For instance, (Kennedy et al., 2020) uses explanations to determine if the model is oversensitive to protected attributes. If so, it applies explanation regularization to ensure the model focuses on the context of these identifiers instead of the group identity terms.

Towards Fair AI

As AI systems are increasingly deployed in critical areas like education, healthcare, and employment, it is essential to address their biases, which can lead to discrimination and harm, to ensure these systems contribute to societal good. The field of fairness-aware learning offers solutions through bias detection, mitigation tools, and proactive methods, which should be integrated into AI pipelines alongside necessary bias-safety checks.

It is important to note that fairness is not a fixed concept but evolves in response to changing societal needs (Rawls, 1971; Sen, 2009). Rapid technological advancements in AI, such as the emerging bias and discrimination issues in generative AI (Hacker et al., 2024), further complicate fairness considerations. Similarly, the evolving legal landscape, such as the AI Act[4] adds additional layers of complexity. Given that fairness is a moving target, continuous monitoring and ongoing adjustments, both during system design and after deployment, are essential to ensure AI systems remain aligned with these shifting societal, technological, and legal expectations.

6 Conclusions

While biases are often perceived negatively for leading to unfair outcomes and discrimination, certain types of biases are crucial for induction and effective learning. In this work, we aimed to provide a broader perspective on biases in AI/ML models, advocating for a balanced view of bias as both a challenge and a tool. Biases that lead to discrimination and harm should be identified and mitigated, while others, such as inductive biases, are fundamental to the learning process.

Ultimately, understanding these biases and applying the right tools to either mitigate them or leverage them to steer models in the "right direction" is crucial for responsible and effective AI development, ensuring these systems contribute positively to societal good.

Acknowledgement This work was supported by the European Union Horizon Europe Projects: STELAR (Grant Agreement ID: 101070122) and MAMMOth (Grant Agreement ID: 101070285).

[4] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

References

Adewumi, T., Alkhaled, L., Gurung, N., van Boven, G., & Pagliai, I. (2024). Fairness and bias in multimodal AI: A survey. *arXiv, arXiv*, 2406.19097.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. *Proceedings of the International Conference on Machine Learning*, 60–69.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). *Machine bias. Ethics of data and analytics* (S. 254–264). Auerbach Publications.

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 1–41.

Berendt, B., Gandon, F., Halford, S., Hall, W., Hendler, J., Kinder-Kurlanda, K., Ntoutsi, E., & Staab, S. (2021). *Web futures: Inclusive, intelligent, sustainable. The 2020 Manifesto for Web Science*.

Blodgett, S. L., Barocas, S., Daumé, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of bias in NLP. *arXiv, arXiv*, 2005.14050.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29.

Brzezinski, D., Stachowiak, J., Stefanowski, J., Szczech, I., Susmaga, R., Aksenyuk, S., Ivashka, U., & Yasinskyi, O. (2024). Properties of fairness measures in the context of varying class imbalance and protected group ratios. *ACM Transactions on Knowledge Discovery from Data*.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*, 77–91.

Cai, Y., Zimek, A., Wunder, G., & Ntoutsi, E. (2022). Power of explanations: Towards automatic debiasing in hate speech detection. *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 1–10.

Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery, 21*, 277–292.

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building classifiers with independency constraints. *IEEE International Conference on Data Mining Workshops*, 13–18.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems*.

Chhabra, A., Masalkovaite, K., & Mohapatra, P. (2021). An overview of fairness in clustering. *IEEE Access, 9*, 130698–130720.

Dastin, J. (2018, May 12). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Available from: https://tinyurl.com/2vz62ye6. Accessed 12 May 2022.

Deck, L., Schoeffer, J., De-Arteaga, M., & Kühl, N. (2023). A critical survey on fairness benefits of XAI. *arXiv, arXiv*, 2310.13007.

Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery, 3*, 409–425.

Dutta, S., Wei, D., Yueksel, H., Chen, P. Y., Liu, S., & Varshney, K. (2020). Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. *Proceedings of the International Conference on Machine Learning*, 2803–2813.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. S. (2012). Fairness through awareness. *Proceedings of the Conference on Innovations in Theoretical Computer Science*, 214–226.

Dwork, C., Immorlica, N., Kalai, A. T., & Leiserson, M. (2018). Decoupled classifiers for group-fair and efficient machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*, 119–133.

Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., & Kompatsiaris, I. (2022). A survey on bias in visual datasets. *Computer Vision and Image Understanding, 223*, 103552.

Fabris, A., Baranowska, N., Dennis, M. J., Graus, D., Hacker, P., Saldivar, J., et al. (2024). Fairness and bias in algorithmic hiring: A multidisciplinary survey. *ACM Transactions on Intelligent Systems and Technology*.

Fragkathoulas, C., Papanikou, V., Pla Karidi, D., & Pitoura, E. (2024). On explaining unfairness: An overview. *Proceedings of the IEEE International Conference on Data Engineering Workshops*, 226–236.

Fredman, S. (2022). *Discrimination law*. Oxford University Press.

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.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., et al. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, 1–79.

Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé, H., III, et al. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A, 478*(2266), 20210068.

Gupta, J. A. B., Smith, N., & Mariet, Z. (2022). Ensembles of classifiers: A bias-variance perspective. *Transactions on Machine Learning Research*.

Hacker, P., Mittelstadt, B., Zuiderveen Borgesius, F., & Wachter, S. (2024). Generative discrimination: What happens when generative AI exhibits bias, and what can be done about it. *arXiv, arXiv*, 2407.10329.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 29.

Haselton, M. G., Nettle, D., & Andrews, P. W. (2015). The evolution of cognitive bias. In D. M. Buss (Hrsg.), *The handbook of evolutionary psychology* (S. 724–746). Wiley.

Hu, T., Iosifidis, V., Liao, W., Zhang, H., Yang, M. Y., Ntoutsi, E., et al. (2020). FairNN: Conjoint learning of fair representations for fair decisions. In A. Appice, G. Tsoumakas, Y. Manolopoulos, & S. Matwin (Hrsg.), *Discovery Science. DS 2020. Lecture Notes in Computer Science* (Bd. 12323, S. 581–595). Springer.

Hüllermeier, E., Fober, T., & Mernberger, M. (2013). Inductive bias. In W. Dubitzky, O. Wolkenhauer, K. H. Cho, & H. Yokota (Hrsg.), *Encyclopedia of systems biology* (S. 1018). Springer.

Hutchinson, B., & Mitchell, M. (2019). 50 years of test (un)fairness: Lessons for machine learning. *Conference on Fairness, Accountability, and Transparency (FAccT)*, 49–58.

Iosifidis, V., & Ntoutsi, E. (2018). Dealing with bias via data augmentation in supervised learning scenarios. *Proceedings of the International Workshop on Bias in Information, Algorithms, and Systems*.

Iosifidis, V., & Ntoutsi, E. (2019). AdaFair: Cumulative fairness adaptive boosting. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (S. 781–790).

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*(5), 429–449.

Jiang, H., & Nachum, O. (2020). Identifying and correcting label bias in machine learning. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 702–712.

Jones, C. (2020). Law enforcement use of facial recognition: Bias, disparate impacts on people of color, and the need for federal legislation. *North Carolina Journal of Law & Technology, 22*, 777.

Kamiran, F., & Calders, T. (2009). Classifying without discriminating. *IEEE International Conference on Computer, Control, and Communication*, 1–6.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. *IEEE International Conference on Data Mining (ICDM)*, 869–874.

Kamiran, F., Mansha, S., Karim, A., & Zhang, X. (2018). Exploiting reject option in classification for social discrimination control. *Information Sciences, 425*, 18–33.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 35–50.

Kelly, T. (2022). *Bias: A philosophical study*. Oxford University Press.

Kennedy, B., Jin, X., Mostafazadeh Davani, A., Dehghani, M., & Ren, X. (2020). Contextualizing hate speech classifiers with post-hoc explanation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).*

Kim, Y., Mo, S., Kim, M., Lee, K., Lee, J., & Shin, J. (2024). Discovering and mitigating visual biases through keyword explanation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11082–11092.

Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., & Kompatsiaris, Y. (2018). Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. *Proceedings of the World Wide Web Conference (WWW)*, 853–862.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Langley, P., & Sage, S. (1994). Induction of selective Bayesian classifiers. *Uncertainty in Artificial Intelligence*, 399–406.

Luong, B. T., Ruggieri, S., & Turini, F. (2011). k-NN as an implementation of situation testing for discrimination discovery and prevention. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 502–510.

Maity, S., Mukherjee, D., Yurochkin, M., & Sun, Y. (2020). There is no trade-off: Enforcing fairness can improve accuracy. *Stat, 1050*, 6.

Mehlhorn, K. (2013). *Data structures and algorithms 1: Sorting and searching* (Bd. 1). Springer Science & Business Media.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*, 220–229.

Mitchell, T. M. (1980). *The need for biases in learning generalizations.*

Mitchell, T. M. (1997). *Machine learning.*

Montañez, G. D., Hayase, J., Lauw, J., Macias, D., Trikha, A., & Vendemiatti, J. (2019). The futility of bias-free learning and search. *Australasian Joint Conference on Artificial Intelligence*, 277–288.

Murray, K., & Chiang, D. (2018). Correcting length bias in neural machine translation. *arXiv preprint, arXiv*, 1808.10006.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., Ruggieri, S., Turini, F., Papadopoulos, S., & Krasanakis, E. (2020). Bias in data-driven artificial intelligence systems – An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1356.

Padala, M., & Gujar, S. (2020). FNNc: Achieving fairness through neural networks. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI).*

Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 560–568.

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2019). Meaningful explanations of black box AI decision systems. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*, 9780–9784.

Prince, A. E., & Schwarcz, D. (2019). Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review, 105*, 1257.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Roy, A., & Ntoutsi, E. (2022). Learning to teach fairness-aware deep multi-task learning. *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 710–726.

Roy, A., Iosifidis, V., & Ntoutsi, E. (2022). Multi-fairness under class-imbalance. *International Conference on Discovery Science (DS)*, 286–301.

Roy, A., Horstmann, J., & Ntoutsi, E. (2023). Multi-dimensional discrimination in law and machine learning – A comparative overview. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 89–100.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach* (3. Aufl.). Pearson.

Russo, M., & Vidal, M. E. (2024). Leveraging ontologies to document bias in data. *arXiv preprint, arXiv*, 2407.00509.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Schneider, M., Schelte, T., Schmitz, F., & Körner, M. (2023). Eurocrops: The largest harmonized open crop dataset across the European Union. *Scientific Data, 10*(1), 612.

Sen, A. (2009). *The idea of justice*. Penguin Books.

Srinivasan, R., & Chander, A. (2021). Biases in AI systems. *Communications of the ACM, 64*(8), 44–49.

Tian, Y., & Zhang, Y. (2022). A comprehensive survey on regularization strategies in machine learning. *Information Fusion, 80*, 146–166.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *International Workshop on Software Fairness*, 1–7.

Wang, Z., & Wu, L. (2024). Theoretical analysis of the inductive biases in deep convolutional networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Weber, L., Lapuschkin, S., Binder, A., & Samek, W. (2023). Beyond explaining: Opportunities and challenges of XAI-based model improvement. *Information Fusion, 92*, 154–176.

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems. *AI Now*, 1–33.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., & Schwartz, O. (2018). *AI now report 2018*. AI Now Institute at New York University.

Yeom, S., Datta, A., & Fredrikson, M. (2018). Hunting for discriminatory proxies in linear regression models. *Advances in Neural Information Processing Systems (NeurIPS)*, 31.

Yu, R., Lee, H., & Kizilcec, R. F. (2021). Should college dropout prediction models include protected attributes? *Proceedings of the ACM Conference on Learning@ Scale*, 91–100.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *Proceedings of the International World Wide Web Conference (WWW)*, 1171–1180.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 335–340.

Zhang, W., & Ntoutsi, E. (2019). FAHT: An adaptive fairness-aware decision tree classifier. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 1480–1486.

Prof. Dr. Eirini Ntoutsi is a professor for Open Source Intelligence at the University of the Bundeswehr Munich since August 2022. Previously, she was a professor at the Free University Berlin and Leibniz University of Hannover. She earned her Ph.D. from the University of Piraeus, Greece, and holds degrees in Computer Engineering and Informatics from the University of Patras. Her research focuses on AI and ML, creating systems that benefit society by promoting fairness, explainability, and responsibility. An active member of the community, she co-chaired the research track of ECML PKDD 2024 and is recipient of multiple awards, including the Humboldt Fellowship.