Neurosymbolic Architectures for Algorithmic Fairness

Neurosymbolic Artificial Intelligence
XX(X):2–32

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Leonhard Kestel[12], Christoph Kern[12]

Abstract

Bias is a pervasive issue in Machine Learning, particularly in domains like *automated decision-making* (ADM), where it can lead to unfair treatment of individuals or groups based on sensitive attributes. Accounting for it requires knowledge and reasoning about how bias can affect the decision process and how to constrain this process in order to decrease its vulnerability to societal and statistical bias. In the field of bias mitigation, a broad set of constraining techniques has been developed to address the issue of biased predictions. Usually, such a technique is an architecture or procedure particularly designed for a use case or a distinct definition of fairness. In application however, practitioners face complex realities requiring flexible, complex reasoning about constraints, yet the link to integrative approaches that combine symbolic reasoning with logical constraints and statistical learning is still missing. Although there exist several neurosymbolic architectures able to incorporate knowledge and constraints into a model, only few attempts have been made to use them to apply fairness constraints to model predictions. This work tries to bridge this gap by mapping neurosymbolic architectures to bias mitigation techniques. We categorize these architectures based on their potential application in pre-processing, in-processing, and post-processing. By doing so, we aim to provide a structured overview of the current set of existing neurosymbolic architectures for bias mitigation, and highlight important underexplored directions and promising research avenues at the intersection of neurosymbolic AI and algorithmic fairness.

Keywords

Neurosymbolic AI, Algorithmic Fairness, Bias Mitigation, Trustworthy AI

Prepared using sagej.cls [Version: 2017/01/17 v1.20]

1 Introduction

Machine learning models are becoming more and more omnipresent as algorithmic decision makers in various fields, such as public policy making, healthcare and hiring. These domains, in which decisions directly concern the life of people, crucially require trustworthy models. In the context of machine learning, trustworthiness comprises aspects such as *interpretability* (the decision process is understandable), *accountability* (the decision process is underlying clear responsibilities and strict governance), *fairness* (the decision process is not systematically discriminating people), *robustness* (the decision process is not vulnerable to data shifts and data poisoning), *safety* (the decision does not endanger anybody) or *privacy* (the decision does not provide any insight about a person). Fairness stands out among these concepts, as this is an aspect of trustworthiness that ADM actually promises to enhance compared to human decisions. Instead of the bias of the human decision maker, algorithmic data-driven decisions are prone to data bias. Therefore, the field of fair machine learning concerns itself with how to detect and mitigate bias. While bias detection queries whether data or a prediction satisfies a fairness constraint, bias mitigation employs fairness constraints on the data, the prediction model or the output. Among the various different approaches to bias mitigation, many techniques are catering a specific fairness notion, i.e., are tied to one single formal definition of fairness (Caton and Haas 2024; Hort et al. 2022).

Fairness in Automated Decision Making. In ADM practice, e.g., in public policy settings, trustworthiness and fairness in particular are a complex and evolving issue, with nuanced requirements that may change over time. The desired ADM system in this domain is supposed to reliably support decision-making, while operating in an area which often includes multiple stakeholders, competing policy goals, dynamic data streams, as well as complex sources of data biases next to specific regulatory constraints (Abaigar et al. 2024). Furthermore, implementing an algorithmic system in administrative practice commonly requires considerable (time) investments and institutional resources, e.g., in terms of building the technological infrastructure and the training of staff (Wirtz et al. 2019). In such settings, *flexible* and *transparent* approaches to algorithmic fairness are critical, as switching between different modeling

[1]Ludwig-Maximilians-Universität München, Germany
[2]Munich Center for Machine Learning (MCML)

Corresponding author:
Leonhard Kestel
LMU München
Department of Statistics
Ludwigstraße 33 | 80539 Munich, Germany.
Email: leo.kestel@lmu.de

Prepared using sagej.cls

procedures once a system is in place can incur considerable costs and institutional overhead.

In order to achieve this flexibility and transparency, an interface between statistical (neural) models on the one side and a declarative formalization language for symbolic constraints is required. Researchers in the field of neurosymbolic AI have proposed numerous architectures that incorporate the understandable, reasonable nature of symbols and statistical models that can handle noise and uncertainty. Neurosymbolic models hence provide this *flexible interface* between formalized *declarative* constraints and their implementation into the machine learning procedure. Most bias mitigation approaches encode constraints directly into the machine learning *procedure* and thus implement specific fairness notions for a distinct set of use cases. In contrast, a well-designed neurosymbolic bias mitigation method promises to be agnostic regarding the definition of fairness and the bias model, i.e. assumptions about variable dependencies, of the domain. In summary, by concretizing normative concerns into formal rules that a prediction system should adhere to, the field of algorithmic fairness naturally lends itself to the integration of symbolic reasoning and can strongly benefit from the rich set of architectures proposed in neurosymbolic AI.

Another interesting aspect brought on the table by the flexibility of neurosymbolic models, is the opportunity to easily compare the behavior of a predictor under varying constraints. This is important for ADM practitioners, since usually there is no clear case for one distinct fairness notion or one bias model (e.g. Chouldechova 2017; Mitchell et al. 2021). Hence, being able to experiment with different notions and assumptions pre-deployment, e.g. by adding or removing a logical constraint, is desirable.

Symbolic Reasoning and Neural Inference. Symbolic reasoning algorithms process symbols, i.e. discrete meaningful units. Symbolic models usually consists of a knowledge base containing formalized facts and a solver to perform deduction, which is called reasoning. Thus, they are inherently interpretable as their processes, as well as all data representations are explicit and interpretable. The biggest issue in symbolic systems is the grounding problem, i.e. to find an adequate mapping between the continuous real world and the assumed discrete world of the model.

Neural models are complex arithmetic functions with many parameters that process continuous data, transforming it to latent intermediate representations. They require (almost) no prior knowledge as they perform induction on the data, which is optimized according to a loss function. The parameters, which are optimized during the training process represent implicit knowledge that is not interpretable for humans. Hence, their biggest issues comprise interpretability and other aspects of trustworthiness, such as accountability and fairness. Another weakness of neural systems is complex reasoning.

The integration of these two worlds can be seen as an approach to enhance trustworthiness and complex reasoning abilities of neural models or as a promising approach to symbol grounding and the integration of latent/implicit subsymbolic knowledge.

Contribution. Michel-Delétie and Sarker (2024) argue that research on neurosymbolic trustworthy models is focused on leveraging symbolic properties for interpretability and robustness, while lacking on fairness and privacy. Especially regarding fairness, they point towards untapped potential, while stating that neurosymbolic approaches are often not flagged as such. To our knowledge, there are only three approaches, which explicitly use neurosymbolic models for fairness (Wagner and d'Avila Garcez 2021; Greco et al. 2023; Heilmann et al. 2025). All of them emphasize the potential of neurosymbolic AI as a flexible, generalized approach to bias mitigation.

In this work, we aim to systematically bridge the gap between these two domains. Thereby, we point at existing implicitly neurosymbolic approaches to bias mitigation, and propose conceptual architectures, which we find interesting for future research. The incentives of this work are to

1. deliver an argument for the integration of neural and symbolic systems in ethical and fair AI.
2. connect two almost completely disjoint domains and inspire interdisciplinary research.
3. give an overview over existing (neurosymbolic) fairness and bias mitigation research
4. propose a set of neurosymbolic architectures for bias mitigation for future research.

2 Algorithmic Fairness

Fairness in machine learning is a complex multidisciplinary topic that has been studied from various perspectives, including computer science, ethics, law, and social sciences (Baumann and Rumberger 2018). In the following, we provide a brief overview of the most common definitions of fairness in machine learning, as well as a summary of the discussion about these definitions and their associated metrics. For a more comprehensive overview, we refer to surveys by Caton and Haas (2024), Hort et al. (2022) or Mitchell et al. (2021).

Many wide-spread notions of fairness focus on binary classification tasks with one binary protected attribute. Protected attributes may identify different demographic groups as defined in anti-discrimination law (Simson et al. 2024), but can also refer to ascribed or socially constructed characteristics more broadly. Next to group fairness for binary classification, there are also definitions for multiclass classification, regression tasks and multiple, as well as many-valued, protected attributes. In the following, we try to generalize the different definitions as good as possible over a broad spectrum of data and predictors.

2.1 Group Fairness

Group fairness notions require that certain statistical measures of the predictor's performance are equal across different demographic groups A defined by a set of protected attributes. Common (types of) group fairness definitions include e.g.:

Independence. The prediction $\hat{Y}$ is independent of the demographic group. This concept is also known as demographic parity or statistical parity. Independence means that the predictions are distributed equally across groups.

$$P(\hat{Y}|A = a) = P(\hat{Y}|A = a') \quad \forall a, a' \in A \tag{1}$$

Equality of Accuracy. The accuracy of the predictor $\hat{Y}$ is independent of the demographic group A. This means that the predictor has the same accuracy across different demographic groups.

$$P(\hat{Y} = Y|A = a) = P(\hat{Y} = Y|A = a') \quad \forall a, a' \in A \tag{2}$$

Separation. The prediction $\hat{Y}$ and the demographic group A are independent, given the true label Y. This is also known as equalized odds (Hardt et al. 2016). Separation means that the predictor has the same error rates across different demographic groups.

$$P(\hat{Y}|Y = y, A = a) = P(\hat{Y}|Y = y, A = a') \quad \forall a, b \in A, y \in Y \tag{3}$$

Sufficiency. The true label Y and the demographic group A are independent, given the prediction $\hat{Y}$. This is also known as predictive parity (Chouldechova 2017). Sufficiency means that the predictions have the same informative value across different demographic groups.

$$P(Y|\hat{Y} = \hat{y}, A = a) = P(Y|\hat{Y} = \hat{y}, A = a') \quad \forall a, b \in A, \hat{y} \in \hat{Y} \tag{4}$$

Analogously to the precision-recall trade-off, sufficiency and separation have been shown to be mutually exclusive, except in the case of perfect prediction or if the demographic group is independent of the true label (Chouldechova 2017)

2.2 Multi-Group Fairness

Multi-group fairness notions strike a balance between group and individual fairness by extending group-based fairness definitions to larger collections of subgroups and their intersections. Next to rich subgroup fairness Kearns et al. (2018), multi-calibration Hébert-Johnson et al. (2018) and multi-accuracy Kim et al. (2019) represent prominent types of multi-group fairness notions. For a given distribution $\mathcal{D}$ and class of functions $\mathcal{C}$, multi-accuracy requires that the predicted scores of a predictor $\tilde{p} : \mathcal{X} \to [0, 1]$ are unbiased (up to α) across every subpopulation defined by $c \in \mathcal{C}$:

$$\left|\mathbf{E}_{(X,Y)\sim\mathcal{D}}\left[\, c(X) \cdot (Y - \tilde{p}(X)) \,\right]\right| \leq \alpha \tag{5}$$

In contrast to group fairness, the subpopulations may be defined by arbitrary complex combinations of attributes using a function of class $\mathcal{C}$, and are not restricted to discrete classes. Multi-calibration provides a stronger version of multi-accuracy by requiring calibration (rather than just unbiasedness) across collections of subpopulations. Both notions have been studied in various contexts and have proven their value in settings beyond algorithmic fairness (Dwork et al. 2021; Gopalan et al. 2022; Kim et al. 2022; Gopalan et al. 2023b,a; Kern et al. 2024; Pfisterer et al. 2021).

2.3 Individual Fairness

Individual fairness notions require that similar individuals are treated similarly by the predictor, while different individuals are treated differently. Mitchell et al. (2021) call this definition *metric fairness*. One of the most common definitions proposed by of individual fairness Dwork et al. (2012) is the Lipschitz condition.

$$\exists k \in \mathbb{R} : \forall (x_1, \hat{y}_1), (x_2, \hat{y}_2) \in (X, \hat{Y}) : \mathsf{d}_X(x_1, x_2) \leq k \cdot \mathsf{d}_Y(\hat{y}_1, \hat{y}_2) \tag{6}$$

This definition requires a suitable distance function on the input space X and the output space Y. However, finding such a distance function is often difficult and requires domain knowledge. Furthermore, it requires a suitable scaling factor k, which is usually unknown. Finding k is usually implemented as a minimization problem (Dwork et al. 2012).

2.4 Causal Fairness

Causal fairness notions require that the predictor is not influenced by the protected attribute A or any of its descendants in a causal graph. This means that the predictor should not be affected by any causal path from the protected attribute to the prediction. The most common way to formalize causal fairness is through counterfactuals, i.e., what would the prediction be if the demographic group were different, but everything else remained the same. Kusner et al. (2017) proposed counterfactual fairness in different variants:

Individual Counterfactual Fairness. The outcome $\hat{Y}$ of a predictor should be the same in the actual world X as in a counterfactual world X', in which the individual belongs to a different demographic group. This notion is similar to individual fairness, but instead of a distance function, it is based on causal interventions.

$$\hat{y}_x = \hat{y}_{x'} \quad \forall (x, x') \in (X, X') \tag{7}$$

Counterfactual Parity. The distribution of the predictor's outcome should be the same in the actual world as in a counterfactual world, in which the individual belongs to a different demographic group. The notion is quite similar to independence, but it is based on causal interventions instead of statistical measures.

$$P(\hat{Y}|X) = P(\hat{Y}|X') \tag{8}$$

3 Bias Mitigation

In this section, we give a condensed, yet structured, overview over bias mitigation techniques, based on comprehensive surveys by Hort et al. (2022); Caton and Haas (2024). With rising concerns about algorithmic fairness in the last two decades, a collection of techniques to reduce bias in machine learning inference has been developed. These can be roughly categorized by the stage of the learning process, they are applied in: *pre-processing* (before training), *in-processing* (during training), and *post-processing* (after training). In many cases however, this categorization is ambiguous, as some methods are applied during multiple stages of the process. Also, these techniques are not exclusive, but can be applied in combination.

3.1 Pre-Processing

Pre-processing techniques aim to remove the bias from the training data, assuming that a predictor trained on fair data is fair. Usually, they come with the advantage that they are model agnostic, as they are mainly concerned with data. Furthermore, Akintande et al. (2025) argue that bias mitigation at a later stage is vulnerable against systematic label bias. Kusner et al. (2017) add to that argument by stating that a model trained on the ideal dataset with perfect accuracy will satisfy independence, separation, calibration, and counterfactual fairness.

We try to classify pre-processing techniques by the data dimension (feature space or instance space) they apply manipulations to and add a third family that constructs a mapping towards a latent fair representation of the entire data. Finally, we discuss fair data generation approaches.

Feature Manipulation. In this family of pre-processing techniques, ground truth labels (relabeling) or predictive feature values (perturbation) are adjusted. For binary relabeling, *massaging* is an established method (e.g. Kamiran and Calders 2009; Calders et al. 2009; Zliobaite et al. 2011), which ranks instances and flips those closest to the decision boundary. This method is based on the assumption that unprivileged individuals scratch this boundary from below. Multiclass relabeling and perturbation of confounded features are often realized by a causal intervention on the feature distribution (e.g. Feldman et al. 2015; Bothmann et al. 2023). Another approach by Lum and Johndrow (2016) proposes a transformation to achieve independence between any feature and the protected attributes.

Instead of manipulating feature values, some researchers introduced latent variables as balanced proxies for labels (e.g. Chakraborty et al. 2022; Calders and Verwer 2010) or group memberships (e.g. Diana et al. 2022; Oneto et al. 2019; Suriyakumar et al. 2023) that follow a given fairness constraint. Similarly, causal inference researchers estimate latent variables as unobserved confounders in their model (Grari et al. 2022; Kilbertus et al. 2017; Madras et al. 2019).

Finally, there is some literature on dropping sensitive and/or proxy features (e.g. Grgic-Hlaca et al. 2018; Madhavan and Wadhwa 2020; Wang and Huang 2019).

Instance Manipulation. Instances can be reweighed, to reduce the impact of potentially biased - and increase the impact of unbiased data points (e.g. Calders et al. 2009; Chai and Wang 2022; Li and Liu 2022), or sampled to reduce misrepresentation of protected groups. The latter can be realized as *downsampling*, i.e. removal of instances (e.g. Chakraborty et al. 2020; Salimi et al. 2019; Wang et al. 2022), *upsampling*, i.e. duplication or synthetization of instances (e.g. Chakraborty et al. 2021; Amend and Spurlock 2021; Chakraborty et al. 2022), *preferential sampling*, i.e. duplication or removal of instances close to the decision border (e.g. Kamiran and Calders 2011; Hu et al. 2020; Zliobaite et al. 2011). Sharma et al. (2020) supplemented data by sampling counterfactual instances using a "realism function" regarding the original data. Abusitta et al. (2019) used a *Generative Adverserial Networks* approach (GAN) to synthesize additional instances for each population group.

Latent Representation. A yet different approach than instance - or feature manipulation is the transformation of an original dataset into an intermediate/latent representation that satisfies fairness constraints and yet retains (almost) all information of the dataset. Starting from a framework called *Learning Fair Representations* (Zemel et al. 2013), many studies have been conducted around this approach, e.g., using optimization (e.g. Calmon et al. 2017; Lahoti et al. 2019; Zehlike et al. 2020), adversarial learning (e.g. Madras et al. 2018a; Qi et al. 2022; Wu et al. 2022), dimensionality reduction (Kamani et al. 2022; Pérez-Suay et al. 2017; Samadi et al. 2018) or with variational autoencoders (e.g. Creager et al. 2019; Liu et al. 2023; Wu et al. 2022)

Data Generation. Starting from a non-neural algorithm (Zhang et al. 2017), Xu et al. (2018, 2019a,b) developed a GAN-based framework, in which datasets are generated from scratch, while one adversary is trained to discriminate fake from real data and another is trained to guess a protected attribute. Other groups have proposed similar GAN-based approaches to fair data generation (Jang et al. 2021; Rajabi and Garibay 2022).

Robertson et al. (2025) modeled *Structured Causal Models* (SCM), representing different types causal influence of protected attributes, as *Multi Layer Perceptrons* (MLP), in which this causal influence can be controlled by a dropout layer. With these, they created two synthetic datasets, one biased and one counterfactual unbiased version, that are later compared in the loss function. This is a textbook example of counterfactual fairness, in which the causal influence of a protected attribute on other features is modeled and controlled in its entirety.

3.2 In-Processing Techniques

Bias mitigation methods that are applied during training address the model instead of the data. Instead of simulating a fictitious world by adjusting the data, the model is constrained to intrinsically learn unbiased predictions on potentially biased data (Wan et al. 2023). Thus, this family of techniques is beneficial in terms of external validity of a model, as it is trained on real-world data and learns how to handle real-world

bias. Furthermore, this approach provides the practical perk of being applicable to pre-trained models (Wan et al. 2023).

Existing in-processing techniques can be roughly classified by the location of their application: the loss function or the training algorithm.

Fairness-Aware Loss Functions. Given a model that may be trained with gradient descent, a quite straight-forward approach is to add an additive regularization term to the loss function. This means that a discriminatory prediction leads to a higher loss and is thus penalized. Hence, one can optimize a model regarding accuracy as well as a metric based on a distinct notion of fairness, e.g. independence, separation, or the distance between a prediction and its counterfactual counterpart (Robertson et al. 2025; Tavakol 2020).

Other models may be regularized differently, e.g. decision trees can be modified to incorporate fairness metrics as splitting criteria (e.g. Kamiran et al. 2010; Ranzato et al. 2021; Zhang and Weiss 2023).

An uprising approach to loss functions is *adversarial learning* (Dalvi et al. 2004). Here, an adversary model is introduced, which is trained to exploit errors of the main model. Its loss is modeled as a *minimax* function, which the main model wants to minimize, while its adversary aims to maximize. In the fairness context, the adversary usually tries to guess a protected attribute from the prediction of a model (e.g. Beutel et al. 2017; Raff and Sylvester 2018; Sadeghi et al. 2019). This is an operationalization of the group fairness notion of independence.

In the taxonomy by Kautz (2022), there is a distinct class of integration, written as $\text{Neuro}_{\text{Symbolic}}$, that comprises methods, which incorporate symbolic rules into the loss function of a neural network. Widespread frameworks of this class are LTNs (Serafini and d'Avila Garcez 2016) as described in Section 4. As one of the first works on neurosymbolic fairness, Wagner and d'Avila Garcez (2021) proposed an LTN that incorporates group fairness constraints in first-order logic (FOL). Heilmann et al. (2025) extended this approach with the notion of counterfactual fairness.

Fairness-Aware Training Algorithms. A method for accurate predictions that pays tribute to group fairness is *model composition*. A straight-forward way of this is training multiple models for each population subgroup (e.g., privileged and unprivileged) (e.g. Calders and Verwer 2010; Oneto et al. 2019; Pleiss et al. 2017; Suriyakumar et al. 2023). Instead of just picking the outcome of the regarding predictor, predictions can be aggregated in an ensemble fashion, so that multiple models with different fairness or accuracy goals can be taken into account (e.g. Liu and Vicente 2022; Mishler and Kennedy 2022; Valdivia et al. 2021).

Adjusted learning on the other hand, provides a set of techniques to alter or recreate the learning procedure. Usually, these methods look at critical data points with respect to a fairness metric and treat them differently. E.g. Noriega-Campero et al. (2019); Anahideh et al. (2022) used *active learning* methods that query for more information on these data points to retrain them, Madras et al. (2018b) proposed a *rejection learning* approach to learn, when to defer from making a prediction, while Hébert-Johnson et al. (2018) proposed a boosting-like algorithm for multicalibration.

Other research focused on *hyperparameter tuning* (e.g. Chakraborty et al. 2020, 2019; Valdivia et al. 2021).

3.3 Post-Processing Techniques

Post-processing techniques assume a completely trained model that can make biased decisions. They are rather concerned about the (hard) correction of input, model or output towards an unbiased prediction than imposing (soft) constraints on the learning environment of the model. Lohia et al. (2019) argue that post-processing techniques are becoming especially useful nowadays, because the model training and deployment are often decoupled. Hence, a model may be pre-trained by a third party and only accessible as a black-box API. In this case, pre- or in-processing techniques are not applicable.

Input Correction. Hort et al. (2022) argue that input correction uses the same set of techniques as pre-processing, e.g. perturbation (Adler et al. 2018; Li et al. 2022), since it adjusts the input data. However, it is applied to test data instead of training data.

Model Correction. Similar to in-processing techniques, model correction methods adjust the model itself. However, instead of adjusting the initial loss function or learning procedure, they fine-tune or directly manipulate the parameters of successfully trained model. E.g. Savani et al. (2020) proposed three techniques to adjust the weights of a pre-trained neural network to accomodate group fairness metrics: *random weight perturbation*, *layerwise optimization*, and *adversarial fine-tuning*. A much cited and further extended approach by Hardt et al. (2016) uses a linear optimzation algorithm to minimally adjust a classifier to satisfy equality of opportunity or equalized odds. Pleiss et al. (2017) on the other hand split a trained classifier into multiple models for each population subgroup and adjusted their decision boundaries individually to achieve calibration. Kim et al. (2019) proposed a boosting-algorithm to iteratively adjust a model to improve accuracy for certain subgroups.

Output Correction. At the latest stage of the machine learning pipeline, the output can be adjusted. This is often done analogously to the preprocessing approach of relabeling. The selection of instances to be relabeled requires another model, e.g. a ranking of instances close to the decision border (Kamiran et al. 2012, 2018), group-dependent decision thresholds (e.g. Pentyala et al. 2022; Iosifidis et al. 2020), a model optimized to find instances likely to be discriminated (Lohia et al. 2019), or a counterfactual world, which the prediction is aligned to (Chiappa 2019).

4 Neurosymbolic Architectures

Numerous frameworks, which integrate neural inference and symbolic reasoning have already been developed. Kautz (2022) proposed a taxonomy to classify these by the type of their integrations. This taxonomy is now widely used to structure surveys (e.g. Bhuyan et al. 2024; Wan et al. 2024). In the following, we provide a brief overview

over these architecture types. For more details on single approaches, we refer to more comprehensive surveys, e.g. by Bhuyan et al. (2024) and Wan et al. (2024).

Type 1: Symbolic→Neuro→Symbolic. In this architecture type, integration is understood as translation. Symbols are translated to vectors that can be processed by a neural network. The continuous output of this network is finally discretized to symbols again. The classical example for this approach is the current standard approach to NLP that uses word embeddings like *word2vec* (Mikolov et al. 2013) or *glove* (Pennington et al. 2014).

Type 2: Symbolic[Neuro]. This type uses neural models as functions that are called by a symbolic solver. E.g. *AlphaGo* (Silver et al. 2016), the first model to beat a human champion at the game Go, integrated neural heuristics into Monte Carlo Tree Search. In general this method is promising for NP-hard problems, which are solvable by symbolic reasoning, but for which the search space is too large to be explored exhaustively. The neural model can be used as an oracle to prune the search space and thus speed up the search process.

Type 3: Neuro|Symbolic. This architecture type employs neural and symbolic models as coroutines in a pipeline. These coroutines have disjoint equal-levelled tasks E.g. *DeepProbLog* (Manhaeve et al. 2018) uses a neural model to generate a set of candidate rules, which are then evaluated by a symbolic solver. The symbolic solver can then provide feedback to the neural model, which can be used to improve the rule generation process.

Type 4: Neuro:Symbolic→Neuro. A common example for this is the work of Lample and Charton (2020), who proposed a rigorous learning procedure for transformer models to validly transform math formulae. Thus, the neural model itself can perform complex symbolic reasoning steps. In general, this category comprises methods that adjust the weights, the architecture or the learning procedure of a neural model, so that it can do certain symbolic deduction without any other reasoner.

Type 5: Neuro$_{Symbolic}$. This type is becoming more and more popular as a method to add semantic constraints to a loss functions. A common example here are *Logic Tensor Networks* (LTN) (Serafini and d'Avila Garcez 2016), which use fuzzy first-order logic to create differentiable axiomatic loss functions. Neuro$_{\mathbf{Symbolic}}$ techniques are effective at employing soft constraints during the learning process.

Type 6: Neuro[Symbolic]. Based on cognitive dual process theories (e.g. Stanovich and West 2000; Kahneman 2003), this type uses a neural encoder to produce a latent symbolic representation of the data, which a reasoner transforms and forwards to a neural decoder. Conversely to type 2, the symbolic abstraction is embedded in a neural model here. E.g. Asai and Fukunaga (2018) developed a neurosymbolic solver for 8-puzzles that works exactly like this. This category of methods is often emphasized to be the one with the highest potential as it mimics human cognition: a fast, associative system is doing implicit processing that is evaluated, refined, and completed by a slower, explicit system.

5 Neurosymbolic Architectures for Bias Mitigation

In this section, we discuss how the different classes of neurosymbolic architectures by Kautz (2022) can be used for bias mitigation. Along the way, we classify existing approaches and propose potential future directions (see Table 1). Since we currently do not see a use case for type 1 neurosymbolic architectures in bias mitigation, they do not occur further on in this section.

Table 1. Proposal of architectures for bias mitigation. Each row represents a neurosymbolic *architecture* of a distinct *type* (see Section 4), that we propose as a *bias mitigation method* at a specific stage (see Section 3). All of these proposals are elaborated on in Section 5. We introduce the term *Fairness-Aware Model* here as a technique integrating fairness constraints directly into the model architecture

Type	Architecture	Bias Mitigation Method	Reference
S[N]	—	Output Correction	—
N|S	—	Output Correction	—
N:S→N	SCM-based MLP	Data Generation	Hollmann et al. (2023)
N:S→N	LNN	Fairness-Aware Model	—
N:S→N	Differentiable ILP	Fairness-Aware Model; Model Correction	—
N_S	LTN	Fairness-Aware Loss Function	Wagner and d'Avila Garcez (2021); Greco et al. (2023); Heilmann et al. (2025)
N_S	Generative Adversarial LTN	Data Generation	—
N[S]	—	Latent Representation; Data Generation	—

5.1 Symbolic[Neuro] and Neuro|Symbolic Bias Mitigation

The Symbolic[Neuro] and Neuro|Symbolic architectures are quite similar, as they both consist of two distinct independent parts, one neural and one symbolic part. The difference is that in Symbolic[Neuro], the symbolic part is the main driver and the neural part is a subroutine, while in Neuro|Symbolic, the neural part is the main driver and the symbolic part is a subroutine. We will discuss them together, as it is sometimes hard to distinguish between the role of a co- or a subroutine.

As these classes consists of two independent submodels, their architecture is not suitable for any technique that aims to directly alter the prediction model or its learning procedure. However, the symbolic co-routine can be used to relabel the output of a neural model. While naturally being classified as a post-processing method, this approach would be flagged as a pre-processing technique if the output of the model is a generated dataset that is then relabeled.

Chiappa (2019) proposed a method to adjust the output of a predictor to satisfy counterfactual fairness. They use a causal model to generate counterfactuals and adjust the prediction of a model towards its counterfactual counterpart. Though the

reasoning part of this approach is small, it demonstrates the potential of symbolic methods to post-process the output of a neural model.

Such approaches allow for potentially more powerful and precise corrections than e.g. group-dependent thresholds. Furthermore, unlike a statistical bias detector, they provide an interpretable and accountable component: the neural model can make accurate predictions, while the symbolic model ensures that the predictions adhere to prespecified fairness criteria. This approach might be particularly suitable for scenarios where the bias in the data is complex and requires reasoning about multiple attributes.

5.2 $\text{Neuro}_{\text{Symbolic}}$ Bias Mitigation

$\text{Neuro}_{\text{Symbolic}}$ architectures are neural models that incorporate symbolic rules into the loss function of a neural model. Thus, they regularize the learning procedure of a neural network with symbolic axioms, leading to *softly* imposed fairness constraints.

5.2.1 Logic Tensor Networks with Fairness Constraints The probably most prominent $\text{Neuro}_{\text{Symbolic}}$ framework are Logic Tensor Networks (Serafini and d'Avila Garcez 2016) as mentioned in Section 4. They incorporate FOL axioms into the loss function of a neural network by interpreting logical symbols as differentiable fuzzy functions and predicates. Thus, the truth value of a formula can be evaluated in a continuous space and used as a loss term. Wagner and d'Avila Garcez (2021) as well as Greco et al. (2023) used LTNs as a means to include fairness constraints as FOL axioms in the loss of a neural network. Additionally to an accuracy axiom, they used the group fairness metrics *demographic parity* and *disparate impact*. Heilmann et al. (2025) added the notion of *counterfactual fairness* to that approach.

However, much more is possible, as LTNs allow for any constraint that can be formalized in FOL. Consider the following signature*:

$$\Sigma = (\emptyset, \{\mathsf{y}/1, \mathsf{a}/1, \mathsf{cf}/1, \mathsf{d}/2, \mathbf{P}/1, \mathbf{GuessA}/1\}, \{=/2, </2, \bot/2\}) \tag{9}$$

*A signature $\Sigma = (\mathcal{C}, \mathcal{F}, \mathcal{P})$ represents the non-logical symbols (constants $\mathcal{C}$, functions $\mathcal{F}$ and predicates $\mathcal{P}$) of a FOL language.

Table 2. Description of the functions and predicates of the FOL signature for fairness constraints (Equation 9).

Function symbols (neural functions in bold):	
$\mathsf{y}(v)$	(get the ground truth label of a sample v)
$\mathsf{a}(v)$	(get the sensitive attribute of a sample v)
$\mathsf{cf}(v)$	(get the counterfactual of a sample v)
$\mathsf{d}_V(v_1, v_2)$	(distance function, tailored to the variable type V)
$\mathbf{P}(v)$	(prediction on a sample v)
$\mathbf{GuessA}(v)$	(guess the sensitive attribute from a prediction v)
Predicate symbols:	
$u = v$	(infix equality predicate)
$u \leq v$	(infix less-or-equal-than predicate)
$u \perp v$	(u is independent from v: $(\forall v_1, v_2 \in v : \frac{\lvert u \wedge v_1 \rvert}{\lvert v_1 \rvert} = \frac{\lvert u \wedge v_2 \rvert}{\lvert v_2 \rvert})$

Using this signature with a dataset X and the set of possible outcomes Y, we can formalize a variety of fairness constraints as FOL axioms (see also Figure 1). E.g.:

$$\text{Accuracy:} \quad \forall x \in X : \mathbf{P}(x) = \mathsf{y}(x) \tag{10}$$

$$\text{Equality of Accuracy:} \quad \forall x \in X : (\mathbf{P}(x) = \mathsf{y}(x)) \perp \mathsf{a}(x) \tag{11}$$

$$\text{Independence:} \quad \forall x \in X : \mathbf{P}(x) \perp \mathsf{a}(x) \tag{12}$$

$$\text{Separation:} \quad \forall x \in X, y \in Y : \mathsf{y}(x) = y \implies \mathbf{P}(x) \perp \mathsf{a}(x) \tag{13}$$

$$\text{Sufficiency:} \quad \forall x \in X, y \in Y : \mathbf{P}(x) = y \implies \mathsf{y}(x) \perp \mathsf{a}(x) \tag{14}$$

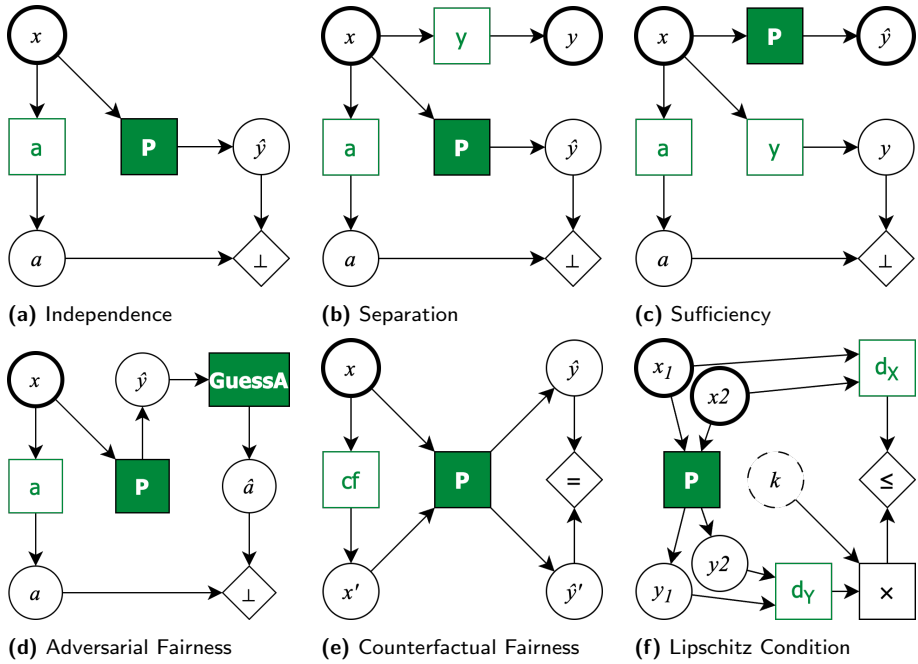
$$\text{Adversarial Fairness:} \quad \forall x \in X : \mathbf{GuessA}(\mathbf{P}(x)) \perp \mathsf{a}(x) \tag{15}$$

$$\text{Counterfactual Fairness:} \quad \forall x \in X : \mathbf{P}(x) = \mathbf{P}(\mathsf{cf}(x)) \tag{16}$$

$$\text{Lipschitz Fairness:} \quad \begin{aligned} &\exists k \in \mathbb{R} : \forall x_1, x_2 \in X : \\ &\mathsf{d}_X(x_1, x_2) \leq k \cdot \mathsf{d}_Y(\mathbf{P}(x_1), \mathbf{P}(x_2)) \end{aligned} \tag{17}$$

Individual fairness, or Lipschitz fairness, resembles a special case, as this axiom performs an existential quantification over an infinite space ($k \in \mathbb{R}$). This is not feasible for symbolic solvers, but can be formulated as an optimization problem as proposed by e.g. Dwork et al. (2012). Hence, incorporating this axiom in an LTN requires a hybrid approach, where the LTN optimizes the neural model to satisfy the other axioms, while another optimization procedure searches for a suitable k.

5.2.2 Generative Adversarial Logic Tensor Networks With the above defined axioms, LTNs can be used as an effective in-processing bias mitigation method. However, a direction that has not yet been explored is the integration LTNs with GANs for fair data generation. The idea behind this integration is to incorporate FOL constraints in addition to the discriminator network into the loss function of a neural data generator. Therefore, some constraints need to be reformulated, after introducing a neural predicate **D** representing the discriminator and the neural function **G** representing the generator, from which we sample datasets. E.g.:



(a) Independence **(b)** Separation **(c)** Sufficiency

(d) Adversarial Fairness **(e)** Counterfactual Fairness **(f)** Lipschitz Condition

Figure 1. Axioms of Algorithmic Fairness: nodes with a bold outline represent universally quantified variables, nodes with a dashed outline represent existentially quantified variables. Functions are represented as rectangles, predicates as diamonds, variables as circles. Neural components are emphasized in green. Each graph uses the signature outlined in Equation 9.

$$\text{Accuracy:} \quad \forall x \in \mathbf{G} : \neg \mathbf{D}(x) \tag{18}$$

$$\text{Equality of Accuracy:} \quad \forall x \in \mathbf{G} : \mathsf{a}(x) \perp (\mathsf{y}(x) = \mathsf{y}(x)) \tag{19}$$

$$\text{Independence:} \quad \forall x \in \mathbf{G} : \mathsf{a}(x) \perp \mathsf{y}(x) \tag{20}$$

$$\text{Adversarial Fairness:} \quad \forall x \in \mathbf{G} : \mathbf{GuessA}(\mathsf{y}(x)) \perp \mathsf{a}(x) \tag{21}$$

$$\text{Counterfactual Fairness:} \quad \forall x \in \mathbf{G} : \mathsf{y}(x) = \mathsf{y}(\mathsf{cf}(x)) \tag{22}$$

$$\text{Lipschitz Fairness:} \quad \begin{array}{l} \exists k \in \mathbb{R} : \forall x_1, x_2 \in \mathbf{G} : \\ \mathsf{d}_X(x_1, x_2) \leq k \cdot \mathsf{d}_Y(\mathsf{y}(x_1), \mathsf{y}(x_2)) \end{array} \tag{23}$$

An important note regarding LTNs also argued by Heilmann et al. (2025) is that they, while optimizing towards fairness constraints, can not guarantee that these constraints are fully satisfied. Instead, they optimize the degree to which these constraints are satisfied. This is a consequence of the fuzzy semantics of LTNs, which allow for a differentiable optimization procedure. Hence, LTNs are best suited for

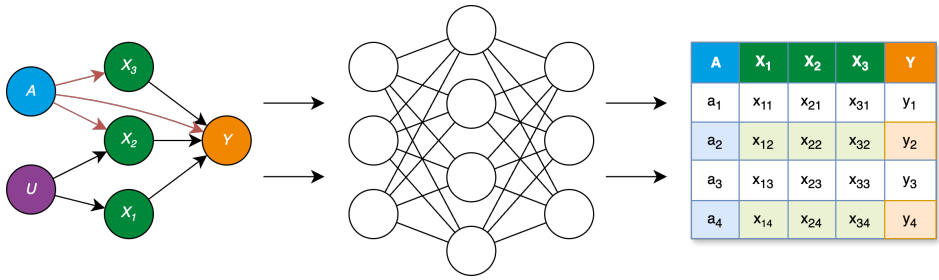


Figure 2. Neuro:Symbolic→Neuro Architecture for SCM-based Data Generation as proposed by Hollmann et al. (2023): An MLP based on an SCM models the causal relationships between features as arithmetic functions. For each instance, a counterfactual version is created by dropping the influence (red arrows) of a sensitive attribute A.

scenarios where approximate fairness is sufficient, but constraints do not need to be satisfied for every single individual.

5.3 Neuro:Symbolic→Neuro Bias Mitigation

Architectures of these class are characterized by a neural model that is rigorously following symbolic rules. Hence, they are particularly suitable for predictions that must satisfy a set of hard constraints for every single prediction. These predictions can be the output of a model or an adjusted or newly generated dataset. Thus, this class of architectures is suitable for both pre- and in-processing bias mitigation techniques. The differentiable ILP framework we discuss in Subsection 5.3.3 is exception, though, since it might be a future direction for post-processing model correction.

5.3.1 SCM Data Generation As mentioned in Section 3.1, Hollmann et al. (2023) developped a method to generate synthetic data using SCMs that are represented as MLPs. Robertson et al. (2025) used this approach to generate alternate versions of a dataset, one with unwanted dependencies and one without. Though never proposed as such, this SCM-based data generation approach is a type 4 neurosymbolic (Neuro:Symbolic→Neuro) method as (neural) MLPs are used as a representation of (symbolic) SCMs to model causal relationships between features as arithmetic functions. What makes this approach unique, is that Robertson et al. (2025) use it to create numerous fair datasets from scratch to feed them to a *TabPFN* model (Hollmann et al. 2023) for pretraining. The TabPFN training process can in turn as well be seen as type 4 neurosymbolic, because a neural foundation model is trained on entirely synthetic datasets, each of which produced by –and thus representing– a structured causal model. Hence, it is rigorously trained to model structured causal relationships. In summary, this approach uses a neurosymbolic data generation approach as a subroutine of a neurosymbolic training procedure.

5.3.2 Logical Neural Networks with Fairness Constraints Logical Neural Networks (LNNs) as introduced by Riegel et al. (2020) are a type 4 neurosymbolic approach

to integrate FOL constraints in the architecture of a neural network. In contrast to LTNs, LNNs do not optimize towards a degree of constraint satisfaction, but instead guarantee that the constraints are fully satisfied (Riegel et al. 2020). This is achieved by a different architecture and semantics. LNNs represent logical formulas as a network of neurons, where each neuron represents a logical connective. The weights of the neurons are constrained to represent the truth tables of the corresponding logical operators. Hence, LNNs can be used to incorporate the same FOL axioms, as formulated in Section 5.2.1, in the architecture of a neural network.

To the best of our knowledge, LNNs have not yet been used for bias mitigation. However, they might be a promising in-processing approach, as they can guarantee that the incorporated fairness constraints are fully satisfied. This makes them particularly suitable for scenarios where a constraint is critical and must be guaranteed for every single individual.

5.3.3 Differentiable ILP to Learn Interpretable Rules Another type 4 neurosymbolic method is the use of differentiable architectures for *Inductive Logic Programming* (ILP) in order to handle noisy and erroneous data. E.g. Evans and Grefenstette (2018) developed such a framework called ∂ILP. Their method learns logical rules from data using a neural network, which makes the ILP procedure less prone to noise. The learned rules can then be used to make predictions on new data. This approach can be used for in- and post-processing bias mitigation by learning rules about relations in data. For once, constraints ca be fed as to the model as background knowledge. Additionally, biased rules can be removed post-hoc by hand or by another model. This approach is particularly interesting, because it learns an interpretable symbolic model of relationships between data attributes that is interpretable and correctable.

However, this method requires a suitable dataset that contains enough information to learn meaningful rules. Evans and Grefenstette (2018) demonstrate the efficacy of their approach on ILP benchmarks, but to the best of our knowledge it has not yet been applied to datasets in fairness-relevant contexts. Furthermore, as in ILP a FOL predicate is learned, differentiable ILP algorithms are limited to binary decisions. Cropper et al. (2022) additionally criticize that ILP systems are non-trivial to handle and that in general, the applicability of these algorithms in real world scenarios is yet to prove.

5.4 Neuro[Symbolic] Bias Mitigation

Neuro[Symbolic] architectures are characterized by a neural model that creates a latent representation, which is then processed by a symbolic model before being decoded by another neural network. This architecture is particularly suitable for scenarios where the input data is high-dimensional and unstructured, e.g. images or text, but the prediction task requires reasoning about structured relationships between attributes. As a toy example, consider a video dataset of numerous actors applying for a role: a neural encoder can create a latent representation of the videos, which is then adjusted by a symbolic model, e.g. an SCM that removes the influence of detected sensitive attributes, before a neural decoder either makes the final prediction

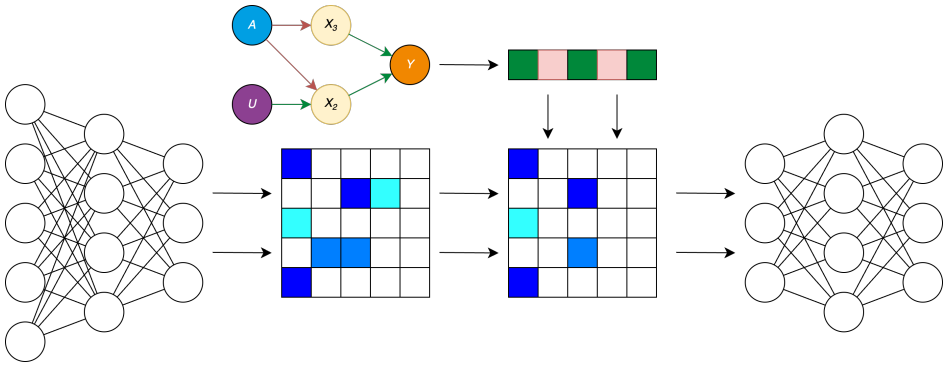


Figure 3. Exemplary Neuro[Symbolic] Architecture: A neural encoder creates a latent representation of relationships between features, which is then processed by layer that represents a structured causal model that eliminates relationships that contain bias (i.e. influences of the protected attribute A) before being processed further. Each column of the matrix in this example represents a relationship between two features as specified by the SCM. Relationships coded in red in the SCM and its vector representation are removed from the matrix.

or recreates a debiased version of the video (see Figure 3). This way, the reasoning model can ensure that the latent representation follows fairness requirements, while the neural model can handle the complexity of the input data and the prediction task. This class of architectures is suitable for both in-processing and pre-processing bias mitigation, as the prediction task can also be data generation.

There are current similar approaches using variational autoencoders with sophisticated optimization procedures to learn fair data representations (e.g. Creager et al. 2019; Liu et al. 2023; Wu et al. 2022). Instead of adjusting the loss function for each new fair approach regarding another fairness notion, these approaches could be extended by logical reasoning. Hence an encoder does not have to be a jack of all trades, but merely learns an intermediate representation that can be interpreted and transformed by a symbolic reasoner. Thus, the same neural model could be used for arbitrary constraints without retraining, while the learning procedure is simpler and the debiasing process is interpretable and controllable.

6 Summary and Propositions

In this work, we provided an overview of neurosymbolic architectures and bias mitigation techniques, and discussed how these two fields can be integrated to create novel bias mitigation methods. Thereby, we argue that different classes of neurosymbolic architectures are suited for different stages of bias mitigation: pre-processing, in-processing, and post-processing. We highlighted existing and potential approaches that utilize neurosymbolic methods for fairness, and discussed their strengths and limitations.

6.1 Claims

Symbolic reasoning provides a means to formalize arbitrary complex constraints. As discussed in Section 5.2.1, many notions proposed in algorithmic fairness can be formalized and composed in FOL. This allows for the integration of these fairness notions into neurosymbolic architectures that support FOL, such as LTNs or LNNs. Another powerful symbolic framework which is widely used are SCMs, which can model causal relationships between features and thus enable the formalization of causal fairness notions, such as counterfactual fairness.

Neurosymbolic AI provides a unifying interface that can accommodate a wide range of fairness notions. Most methods for bias mitigation are designed to address a specific fairness notion, which limits their applicability in scenarios where multiple or alternative notions of fairness are required. Neurosymbolic architectures, on the other hand, provide a flexible framework that can integrate various symbolic representations of fairness notions, allowing for the development of more versatile and adaptable bias mitigation techniques.

Neurosymbolic architectures are a valuable asset on the path towards trustworthy AI. Integrating symbolic reasoning into machine learning models is not only a promising approach for bias mitigation by incorporating constraints, but also for enhancing other aspects of trustworthiness, such as interpretability and robustness. Symbolic reasoning can enhance interpretability by providing clear, rule-based explanations for decisions. Additionally, symbolic constraints can improve robustness by enforcing consistency with a symbolic system. Neurosymbolic model that are more interpretable and controllable can be considered more accountable, as their decisions can be better understood and scrutinized.

These advantages are, however, limited to specific architectures. While e.g. ∂ILP learns an entirely symbolic decision model, other architectures, such as LTNs, use symbolic rules in their training process but provide a black-box neural model at inference time.

Different classes of neurosymbolic architectures are suited for different stages of bias mitigation. As discussed in Section 5, different classes of neurosymbolic architectures have different strengths and weaknesses, which make them more or less suitable for different stages of bias mitigation.

In summary, it is important to note that both *whether* to use one of the proposed architectures and *which* one to use depends on the specific use case and requirements. The choice of architecture should be guided by the nature of the data and task objective, the complexity of the fairness constraints, and the desired level of interpretability and robustness.

6.2 Conclusion

Bias mitigation lacks a generic flexible approach to encoding declarative constraints into the machine learning process. Neurosymbolic models are developed to integrate declarative symbolic knowledge, e.g. constraints, with neural processing.

Our contribution is a first step towards a systematic understanding of how neurosymbolic architectures can be leveraged for bias mitigation in machine learning. By categorizing neurosymbolic architectures and analyzing their applicability to different stages of bias mitigation, we provide an interdisciplinary foundation for researchers and practitioners to explore and develop novel methods that integrate symbolic reasoning with machine learning to address fairness concerns. Thereby, we hope to pave the way for flexible, interpretable and robust methods against machine learning discrimination.

References

Abaigar UF, Kern C, Barda N and Kreuter F (2024) Bridging the gap: Towards an expanded toolkit for ai-driven decision-making in the public sector. *Gov. Inf. Q.* 41(4): 101976. DOI:10.1016/J.GIQ.2024.101976. URL https://doi.org/10.1016/j.giq.2024.101976.

Abusitta A, Aïmeur E and Wahab OA (2019) Generative adversarial networks for mitigating biases in machine learning systems. *CoRR* abs/1905.09972. URL http://arxiv.org/abs/1905.09972.

Adler P, Falk C, Friedler SA, Nix T, Rybeck G, Scheidegger C, Smith B and Venkatasubramanian S (2018) Auditing black-box models for indirect influence. *Knowl. Inf. Syst.* 54(1): 95–122. DOI:10.1007/S10115-017-1116-3. URL https://doi.org/10.1007/s10115-017-1116-3.

Akintande OJ, Bigdeli SA and Feragen A (2025) Medicine after death: XAI and algorithmic fairness under label bias. In: Weerts HJP, Pechenizkiy M, Allhutter D, Corrêa AM, Grote T and Liem CCS (eds.) *European Workshop on Algorithmic Fairness, 30-2 July 2025, Eindhoven University of Technology, Eindhoven, The Netherlands*, *Proceedings of Machine Learning Research*, volume 294. PMLR, pp. 171–186. URL https://proceedings.mlr.press/v294/akintande25a.html.

Amend JJ and Spurlock S (2021) Improving machine learning fairness with sampling and adversarial learning. *J. Comput. Sci. Coll.* 36(5): 14–23. DOI:10.5555/3447307.3447308. URL https://dl.acm.org/doi/10.5555/3447307.3447308.

Anahideh H, Asudeh A and Thirumuruganathan S (2022) Fair active learning. *Expert Syst. Appl.* 199: 116981. DOI:10.1016/J.ESWA.2022.116981. URL https://doi.org/10.1016/j.eswa.2022.116981.

Asai M and Fukunaga A (2018) Classical planning in deep latent space: Bridging the subsymbolic-symbolic boundary. In: McIlraith SA and Weinberger KQ (eds.) *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press, pp. 6094–6101. DOI:10.1609/AAAI.V32I1.12077. URL https://doi.org/10.1609/aaai.v32i1.12077.

Baumann E and Rumberger JL (2018) State of the art in fair ML: from moral philosophy and legislation to fair classifiers. *CoRR* abs/1811.09539. URL http://arxiv.org/abs/1811.09539.

Beutel A, Chen J, Zhao Z and Chi EH (2017) Data decisions and theoretical implications when adversarially learning fair representations. *CoRR* abs/1707.00075. URL http://arxiv.org/abs/1707.00075.

Bhuyan BP, Ramdane-Cherif A, Tomar R and Singh TP (2024) Neuro-symbolic artificial intelligence: a survey. *Neural Comput. Appl.* 36(21): 12809–12844. DOI:10.1007/S00521-024-09960-Z. URL https://doi.org/10.1007/s00521-024-09960-z.

Bothmann L, Dandl S and Schomaker M (2023) Causal fair machine learning via rank-preserving interventional distributions. *CoRR* abs/2307.12797. DOI:10.48550/ARXIV.2307.12797. URL https://doi.org/10.48550/arXiv.2307.12797.

Calders T, Kamiran F and Pechenizkiy M (2009) Building classifiers with independency constraints. In: Saygin Y, Yu JX, Kargupta H, Wang W, Ranka S, Yu PS and Wu X (eds.) *ICDM Workshops 2009, IEEE International Conference on Data Mining Workshops, Miami, Florida, USA, 6 December 2009*. IEEE Computer Society, pp. 13–18. DOI:10.1109/ICDMW.2009.83. URL https://doi.org/10.1109/ICDMW.2009.83.

Calders T and Verwer S (2010) Three naive bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.* 21(2): 277–292. DOI:10.1007/S10618-010-0190-X. URL https://doi.org/10.1007/s10618-010-0190-x.

Calmon FP, Wei D, Vinzamuri B, Ramamurthy KN and Varshney KR (2017) Optimized pre-processing for discrimination prevention. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN and Garnett R (eds.) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. pp. 3992–4001. URL https://proceedings.neurips.cc/paper/2017/hash/9a49a25d845a483fae4be7e341368e36-Abstract.html.

Caton S and Haas C (2024) Fairness in machine learning: A survey. *ACM Comput. Surv.* 56(7): 166:1–166:38. DOI:10.1145/3616865. URL https://doi.org/10.1145/3616865.

Chai J and Wang X (2022) Fairness with adaptive weights. In: Chaudhuri K, Jegelka S, Song L, Szepesvári C, Niu G and Sabato S (eds.) *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, *Proceedings of Machine Learning Research*, volume 162. PMLR, pp. 2853–2866. URL https://proceedings.mlr.press/v162/chai22a.html.

Chakraborty J, Majumder S and Menzies T (2021) Bias in machine learning software: why? how? what to do? In: Spinellis D, Gousios G, Chechik M and Penta MD (eds.) *ESEC/FSE '21: 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Athens, Greece, August 23-28, 2021*. ACM, pp. 429–440. DOI:10.1145/3468264.3468537. URL https://doi.org/10.1145/3468264.3468537.

Chakraborty J, Majumder S and Tu H (2022) Fair-ssl: Building fair ML software with less data. In: *2nd IEEE/ACM International Workshop on Equitable Data & Technology, FairWare@ICSE 2022, Pittsburgh, PA, USA, May 9, 2022*. ACM / IEEE, pp. 1–8. DOI:10.1145/3524491.3527305. URL https://doi.org/10.1145/3524491.3527305.

Chakraborty J, Majumder S, Yu Z and Menzies T (2020) Fairway: a way to build fair ML software. In: Devanbu P, Cohen MB and Zimmermann T (eds.) *ESEC/FSE '20: 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Virtual Event, USA, November 8-13, 2020*. ACM, pp. 654–665. DOI:10.1145/3368089.3409697. URL https://doi.org/10.1145/3368089.3409697.

Chakraborty J, Xia T, Fahid FM and Menzies T (2019) Software engineering for fairness: A case study with hyperparameter optimization. *CoRR* abs/1905.05786. URL http://arxiv.org/abs/1905.05786.

Chiappa S (2019) Path-specific counterfactual fairness. In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, pp. 7801–7808. DOI:10.1609/AAAI.V33I01.33017801. URL https://doi.org/10.1609/aaai.v33i01.33017801.

Chouldechova A (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5(2): 153–163. DOI:10.1089/BIG.2016.0047. URL https://doi.org/10.1089/big.2016.0047.

Creager E, Madras D, Jacobsen J, Weis MA, Swersky K, Pitassi T and Zemel RS (2019) Flexibly fair representation learning by disentanglement. In: Chaudhuri K and Salakhutdinov R (eds.) *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, *Proceedings of Machine Learning Research*, volume 97. PMLR, pp. 1436–1445. URL http://proceedings.mlr.press/v97/creager19a.html.

Cropper A, Dumancic S, Evans R and Muggleton SH (2022) Inductive logic programming at 30. *Mach. Learn.* 111(1): 147–172. DOI:10.1007/S10994-021-06089-1. URL https://doi.org/10.1007/s10994-021-06089-1.

Dalvi NN, Domingos PM, Mausam, Sanghai SK and Verma D (2004) Adversarial classification. In: Kim W, Kohavi R, Gehrke J and DuMouchel W (eds.) *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*. ACM, pp. 99–108. DOI:10.1145/1014052.1014066. URL https://doi.org/10.1145/1014052.1014066.

Diana E, Gill W, Kearns M, Kenthapadi K, Roth A and Sharifi-Malvajerdi S (2022) Multiaccurate proxies for downstream fairness. In: *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM, pp. 1207–1239. DOI:10.1145/3531146.3533180. URL https://doi.org/10.1145/3531146.3533180.

Dwork C, Hardt M, Pitassi T, Reingold O and Zemel RS (2012) Fairness through awareness. In: Goldwasser S (ed.) *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*. ACM, pp. 214–226. DOI:10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

Dwork C, Kim MP, Reingold O, Rothblum GN and Yona G (2021) Outcome indistinguishability. In: Khuller S and Williams VV (eds.) *STOC '21: 53rd Annual ACM*

SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021. ACM, pp. 1095–1108. DOI:10.1145/3406325.3451064. URL https://doi.org/10.1145/3406325.3451064.

Evans R and Grefenstette E (2018) Learning explanatory rules from noisy data. *J. Artif. Intell. Res.* 61: 1–64. DOI:10.1613/JAIR.5714. URL https://doi.org/10.1613/jair.5714.

Feldman M, Friedler SA, Moeller J, Scheidegger C and Venkatasubramanian S (2015) Certifying and removing disparate impact. In: Cao L, Zhang C, Joachims T, Webb GI, Margineantu DD and Williams G (eds.) *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*. ACM, pp. 259–268. DOI:10.1145/2783258.2783311. URL https://doi.org/10.1145/2783258.2783311.

Gopalan P, Hu L, Kim MP, Reingold O and Wieder U (2023a) Loss minimization through the lens of outcome indistinguishability. In: Kalai YT (ed.) *14th Innovations in Theoretical Computer Science Conference, ITCS 2023, January 10-13, 2023, MIT, Cambridge, Massachusetts, USA*, *LIPIcs*, volume 251. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 60:1–60:20. DOI:10.4230/LIPICS.ITCS.2023.60. URL https://doi.org/10.4230/LIPIcs.ITCS.2023.60.

Gopalan P, Kim MP and Reingold O (2023b) Swap agnostic learning, or characterizing omniprediction via multicalibration. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M and Levine S (eds.) *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. URL http://papers.nips.cc/paper_files/paper/2023/hash/7d693203215325902ff9dbdd067a50ac-Abstract-Conference.html.

Gopalan P, Kim MP, Singhal M and Zhao S (2022) Low-degree multicalibration. In: Loh P and Raginsky M (eds.) *Conference on Learning Theory, 2-5 July 2022, London, UK*, *Proceedings of Machine Learning Research*, volume 178. PMLR, pp. 3193–3234. URL https://proceedings.mlr.press/v178/gopalan22a.html.

Grari V, Lamprier S and Detyniecki M (2022) Fairness without the sensitive attribute via causal variational autoencoder. In: Raedt LD (ed.) *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*. ijcai.org, pp. 696–702. DOI:10.24963/IJCAI.2022/98. URL https://doi.org/10.24963/ijcai.2022/98.

Greco G, Alberici F, Palmonari M and Cosentini A (2023) Declarative encoding of fairness in logic tensor networks. In: Gal K, Nowé A, Nalepa GJ, Fairstein R and Radulescu R (eds.) *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, *Frontiers in Artificial Intelligence and Applications*, volume 372. IOS Press, pp. 908–915. DOI:10.3233/FAIA230360. URL https://doi.org/10.3233/FAIA230360.

Grgic-Hlaca N, Zafar MB, Gummadi KP and Weller A (2018) Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In: McIlraith SA and Weinberger KQ (eds.) *Proceedings of the Thirty-Second AAAI*

Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. AAAI Press, pp. 51–60. DOI:10.1609/AAAI.V32I1.11296. URL https://doi.org/10.1609/aaai.v32i1.11296.

Hardt M, Price E and Srebro N (2016) Equality of opportunity in supervised learning. In: Lee DD, Sugiyama M, von Luxburg U, Guyon I and Garnett R (eds.) *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. pp. 3315–3323. URL https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html.

Hébert-Johnson Ú, Kim MP, Reingold O and Rothblum GN (2018) Multicalibration: Calibration for the (computationally-identifiable) masses. In: Dy JG and Krause A (eds.) *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, *Proceedings of Machine Learning Research*, volume 80. PMLR, pp. 1944–1953. URL http://proceedings.mlr.press/v80/hebert-johnson18a.html.

Heilmann X, Manganini C, Cerrato M and Belle V (2025) A neurosymbolic approach to counterfactual fairness. In: *19th International Conference on Neurosymbolic Learning and Reasoning*. URL https://openreview.net/forum?id=YZSDHz3Ydb.

Hollmann N, Müller S, Eggensperger K and Hutter F (2023) Tabpfn: A transformer that solves small tabular classification problems in a second. In: *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. URL https://openreview.net/forum?id=cp5PvcI6w8_.

Hort M, Chen Z, Zhang JM, Sarro F and Harman M (2022) Bias mitigation for machine learning classifiers: A comprehensive survey. *CoRR* abs/2207.07068. DOI:10.48550/ARXIV.2207.07068. URL https://doi.org/10.48550/arXiv.2207.07068.

Hu T, Iosifidis V, Liao W, Zhang H, Yang MY, Ntoutsi E and Rosenhahn B (2020) Fairnn - conjoint learning of fair representations for fair decisions. In: Appice A, Tsoumakas G, Manolopoulos Y and Matwin S (eds.) *Discovery Science - 23rd International Conference, DS 2020, Thessaloniki, Greece, October 19-21, 2020, Proceedings*, *Lecture Notes in Computer Science*, volume 12323. Springer, pp. 581–595. DOI:10.1007/978-3-030-61527-7_38. URL https://doi.org/10.1007/978-3-030-61527-7_38.

Iosifidis V, Fetahu B and Ntoutsi E (2020) FAE: A fairness-aware ensemble framework. *CoRR* abs/2002.00695. URL https://arxiv.org/abs/2002.00695.

Jang T, Zheng F and Wang X (2021) Constructing a fair classifier with generated fair data. In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, pp. 7908–7916. DOI:10.1609/AAAI.V35I9.16965. URL https://doi.org/10.1609/aaai.v35i9.16965.

Kahneman D (2003) Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review* 93(5): 1449–1475. DOI:10.1257/000282803322655392.

URL https://www.aeaweb.org/articles?id=10.1257/000282803322655392.

Kamani MM, Haddadpour F, Forsati R and Mahdavi M (2022) Efficient fair principal component analysis. *Mach. Learn.* 111(10): 3671–3702. DOI:10.1007/S10994-021-06100-9. URL https://doi.org/10.1007/s10994-021-06100-9.

Kamiran F and Calders T (2009) Classifying without discriminating. In: *2009 2nd International Conference on Computer, Control and Communication*. pp. 1–6. DOI:10.1109/IC4.2009.4909197.

Kamiran F and Calders T (2011) Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* 33(1): 1–33. DOI:10.1007/S10115-011-0463-8. URL https://doi.org/10.1007/s10115-011-0463-8.

Kamiran F, Calders T and Pechenizkiy M (2010) Discrimination aware decision tree learning. In: Webb GI, Liu B, Zhang C, Gunopulos D and Wu X (eds.) *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*. IEEE Computer Society, pp. 869–874. DOI:10.1109/ICDM.2010.50. URL https://doi.org/10.1109/ICDM.2010.50.

Kamiran F, Karim A and Zhang X (2012) Decision theory for discrimination-aware classification. In: Zaki MJ, Siebes A, Yu JX, Goethals B, Webb GI and Wu X (eds.) *12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, December 10-13, 2012*. IEEE Computer Society, pp. 924–929. DOI:10.1109/ICDM.2012.45. URL https://doi.org/10.1109/ICDM.2012.45.

Kamiran F, Mansha S, Karim A and Zhang X (2018) Exploiting reject option in classification for social discrimination control. *Inf. Sci.* 425: 18–33. DOI:10.1016/J.INS.2017.09.064. URL https://doi.org/10.1016/j.ins.2017.09.064.

Kautz HA (2022) The third AI summer: AAAI robert s. engelmore memorial lecture. *AI Mag.* 43(1): 93–104. DOI:10.1609/AIMAG.V43I1.19122. URL https://doi.org/10.1609/aimag.v43i1.19122.

Kearns MJ, Neel S, Roth A and Wu ZS (2018) Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: Dy JG and Krause A (eds.) *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, *Proceedings of Machine Learning Research*, volume 80. PMLR, pp. 2569–2577. URL http://proceedings.mlr.press/v80/kearns18a.html.

Kern C, Kim MP and Zhou A (2024) Multi-accurate CATE is robust to unknown covariate shifts. *Trans. Mach. Learn. Res.* 2024. URL https://openreview.net/forum?id=VOGlTb27ob.

Kilbertus N, Rojas-Carulla M, Parascandolo G, Hardt M, Janzing D and Schölkopf B (2017) Avoiding discrimination through causal reasoning. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN and Garnett R (eds.) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. pp. 656–666. URL https://proceedings.neurips.cc/paper/2017/hash/f5f8590cd58a54e94377e6ae2eded4d9-Abstract.html.

Kim MP, Ghorbani A and Zou JY (2019) Multiaccuracy: Black-box post-processing for fairness in classification. In: Conitzer V, Hadfield GK and Vallor S (eds.) *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*. ACM, pp. 247–254. DOI:10.1145/3306618.3314287. URL https://doi.org/10.1145/3306618.3314287.

Kim MP, Kern C, Goldwasser S, Kreuter F and Reingold O (2022) Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences* 119(4). DOI:10.1073/pnas.2108097119.

Kusner MJ, Loftus JR, Russell C and Silva R (2017) Counterfactual fairness. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN and Garnett R (eds.) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. pp. 4066–4076. URL https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

Lahoti P, Gummadi KP and Weikum G (2019) Operationalizing individual fairness with pairwise fair representations. *Proc. VLDB Endow.* 13(4): 506–518. DOI:10.14778/3372716.3372723. URL http://www.vldb.org/pvldb/vol13/p506-lahoti.pdf.

Lample G and Charton F (2020) Deep learning for symbolic mathematics. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL https://openreview.net/forum?id=S1eZYeHFDS.

Li P and Liu H (2022) Achieving fairness at no utility cost via data reweighing with influence. In: Chaudhuri K, Jegelka S, Song L, Szepesvári C, Niu G and Sabato S (eds.) *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, *Proceedings of Machine Learning Research*, volume 162. PMLR, pp. 12917–12930. URL https://proceedings.mlr.press/v162/li22p.html.

Li Y, Meng L, Chen L, Yu L, Wu D, Zhou Y and Xu B (2022) Training data debugging for the fairness of machine learning software. In: *44th IEEE/ACM 44th International Conference on Software Engineering, ICSE 2022, Pittsburgh, PA, USA, May 25-27, 2022*. ACM, pp. 2215–2227. DOI:10.1145/3510003.3510091. URL https://doi.org/10.1145/3510003.3510091.

Liu S, Sun S and Zhao J (2023) Fair transfer learning with factor variational auto-encoder. *Neural Process. Lett.* 55(3): 2049–2061. DOI:10.1007/S11063-022-10920-8. URL https://doi.org/10.1007/s11063-022-10920-8.

Liu S and Vicente LN (2022) Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach. *Comput. Manag. Sci.* 19(3): 513–537. DOI:10.1007/S10287-022-00425-Z. URL https://doi.org/10.1007/s10287-022-00425-z.

Lohia PK, Ramamurthy KN, Bhide M, Saha D, Varshney KR and Puri R (2019) Bias mitigation post-processing for individual and group fairness. In: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. IEEE, pp. 2847–2851. DOI:10.1109/ICASSP.2019.8682620. URL https://doi.org/10.1109/ICASSP.2019.8682620.

Lum K and Johndrow JE (2016) A statistical framework for fair predictive algorithms. *CoRR* abs/1610.08077. URL http://arxiv.org/abs/1610.08077.

Madhavan R and Wadhwa M (2020) Fairness-aware learning with prejudice free representations. In: d'Aquin M, Dietze S, Hauff C, Curry E and Cudré-Mauroux P (eds.) *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. ACM, pp. 2137–2140. DOI: 10.1145/3340531.3412150. URL https://doi.org/10.1145/3340531.3412150.

Madras D, Creager E, Pitassi T and Zemel RS (2018a) Learning adversarially fair and transferable representations. In: Dy JG and Krause A (eds.) *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, *Proceedings of Machine Learning Research*, volume 80. PMLR, pp. 3381–3390. URL http://proceedings.mlr.press/v80/madras18a.html.

Madras D, Creager E, Pitassi T and Zemel RS (2019) Fairness through causal awareness: Learning causal latent-variable models for biased data. In: danah boyd and Morgenstern JH (eds.) *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*. ACM, pp. 349–358. DOI: 10.1145/3287560.3287564. URL https://doi.org/10.1145/3287560.3287564.

Madras D, Pitassi T and Zemel RS (2018b) Predict responsibly: Improving fairness and accuracy by learning to defer. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N and Garnett R (eds.) *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. pp. 6150–6160. URL https://proceedings.neurips.cc/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html.

Manhaeve R, Dumancic S, Kimmig A, Demeester T and Raedt LD (2018) Deepproblog: Neural probabilistic logic programming. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N and Garnett R (eds.) *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. pp. 3753–3763. URL https://proceedings.neurips.cc/paper/2018/hash/dc5d637ed5e62c36ecb73b654b05ba2a-Abstract.html.

Michel-Delétie C and Sarker MK (2024) Neuro-symbolic methods for trustworthy ai: a systematic review. *Neurosymbolic Artificial Intelligence* .

Mikolov T, Chen K, Corrado G and Dean J (2013) Efficient estimation of word representations in vector space. In: Bengio Y and LeCun Y (eds.) *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. URL http://arxiv.org/abs/1301.3781.

Mishler A and Kennedy EH (2022) FADE: fair double ensemble learning for observable and counterfactual outcomes. In: *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM, p. 1053. DOI:10.1145/3531146.3533167. URL https://doi.org/10.1145/3531146.3533167.

Mitchell S, Potash E, Barocas S, D'Amour A and Lum K (2021) Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8(Volume 8, 2021): 141–163. DOI:https://doi.org/10.1146/annurev-statistics-042720-125902. URL https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-042720-125902.

Noriega-Campero A, Bakker MA, Garcia-Bulle B and Pentland AS (2019) Active fairness in algorithmic decision making. In: Conitzer V, Hadfield GK and Vallor S (eds.) *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*. ACM, pp. 77–83. DOI:10.1145/3306618.3314277. URL https://doi.org/10.1145/3306618.3314277.

Oneto L, Donini M, Elders A and Pontil M (2019) Taking advantage of multitask learning for fair classification. In: Conitzer V, Hadfield GK and Vallor S (eds.) *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*. ACM, pp. 227–237. DOI:10.1145/3306618.3314255. URL https://doi.org/10.1145/3306618.3314255.

Pennington J, Socher R and Manning CD (2014) Glove: Global vectors for word representation. In: Moschitti A, Pang B and Daelemans W (eds.) *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, pp. 1532–1543. DOI:10.3115/V1/D14-1162. URL https://doi.org/10.3115/v1/d14-1162.

Pentyala S, Neophytou N, Nascimento ACA, Cock MD and Farnadi G (2022) Privfairfl: Privacy-preserving group fairness in federated learning. *CoRR* abs/2205.11584. DOI:10.48550/ARXIV.2205.11584. URL https://doi.org/10.48550/arXiv.2205.11584.

Pérez-Suay A, Laparra V, Mateo-Garcia G, Muñoz-Marí J, Gómez-Chova L and Camps-Valls G (2017) Fair kernel learning. In: Ceci M, Hollmén J, Todorovski L, Vens C and Dzeroski S (eds.) *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18-22, 2017, Proceedings, Part I*, *Lecture Notes in Computer Science*, volume 10534. Springer, pp. 339–355. DOI:10.1007/978-3-319-71249-9_21. URL https://doi.org/10.1007/978-3-319-71249-9_21.

Pfisterer F, Kern C, Dandl S, Sun M, Kim MP and Bischl B (2021) mcboost: Multi-calibration boosting for R. *J. Open Source Softw.* 6(64): 3453. DOI:10.21105/JOSS.03453. URL https://doi.org/10.21105/joss.03453.

Pleiss G, Raghavan M, Wu F, Kleinberg JM and Weinberger KQ (2017) On fairness and calibration. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN and Garnett R (eds.) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. pp. 5680–5689. URL https://proceedings.neurips.cc/paper/2017/hash/b8b9c74ac526fffbeb2d39ab038d1cd7-Abstract.html.

Qi T, Wu F, Wu C, Lyu L, Xu T, Liao H, Yang Z, Huang Y and Xie X (2022) Fairvfl: A fair vertical federated learning framework with contrastive adversarial learning. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K and Oh A (eds.) *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information*

Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/333a7697dbb67f09249337f81c27d749-Abstract-Conference.html.

Raff E and Sylvester J (2018) Gradient reversal against discrimination: A fair neural network learning approach. In: Bonchi F, Provost FJ, Eliassi-Rad T, Wang W, Cattuto C and Ghani R (eds.) *5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018*. IEEE, pp. 189–198. DOI: 10.1109/DSAA.2018.00029. URL https://doi.org/10.1109/DSAA.2018.00029.

Rajabi A and Garibay ÖÖ (2022) Tabfairgan: Fair tabular data generation with generative adversarial networks. *Mach. Learn. Knowl. Extr.* 4(2): 488–501. DOI:10.3390/MAKE4020022. URL https://doi.org/10.3390/make4020022.

Ranzato F, Urban C and Zanella M (2021) Fairness-aware training of decision trees by abstract interpretation. In: Demartini G, Zuccon G, Culpepper JS, Huang Z and Tong H (eds.) *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*. ACM, pp. 1508–1517. DOI:10.1145/3459637.3482342. URL https://doi.org/10.1145/3459637.3482342.

Riegel R, Gray AG, Luus FPS, Khan N, Makondo N, Akhalwaya IY, Qian H, Fagin R, Barahona F, Sharma U, Ikbal S, Karanam H, Neelam S, Likhyani A and Srivastava SK (2020) Logical neural networks. *CoRR* abs/2006.13155. URL https://arxiv.org/abs/2006.13155.

Robertson J, Hollmann N, Müller S, Awad NH and Hutter F (2025) Fairpfn: A tabular foundation model for causal fairness. *CoRR* abs/2506.07049. DOI:10.48550/ARXIV.2506.07049. URL https://doi.org/10.48550/arXiv.2506.07049.

Sadeghi B, Yu R and Boddeti V (2019) On the global optima of kernelized adversarial representation learning. In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, pp. 7970–7978. DOI:10.1109/ICCV.2019.00806. URL https://doi.org/10.1109/ICCV.2019.00806.

Salimi B, Rodriguez L, Howe B and Suciu D (2019) Interventional fairness: Causal database repair for algorithmic fairness. In: Boncz PA, Manegold S, Ailamaki A, Deshpande A and Kraska T (eds.) *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*. ACM, pp. 793–810. DOI:10.1145/3299869.3319901. URL https://doi.org/10.1145/3299869.3319901.

Samadi S, Tantipongpipat UT, Morgenstern J, Singh M and Vempala SS (2018) The price of fair PCA: one extra dimension. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N and Garnett R (eds.) *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. pp. 10999–11010. URL https://proceedings.neurips.cc/paper/2018/hash/cc4af25fa9d2d5c953496579b75f6f6c-Abstract.html.

Savani Y, White C and Govindarajulu NS (2020) Intra-processing methods for debiasing neural networks. In: Larochelle H, Ranzato M, Hadsell R, Balcan M and Lin H (eds.) *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. URL https://proceedings.neurips.cc/paper/2020/hash/1d8d70dddf147d2d92a634817f01b239-Abstract.html.

Serafini L and d'Avila Garcez AS (2016) Logic tensor networks: Deep learning and logical reasoning from data and knowledge. In: Besold TR, Lamb LC, Serafini L and Tabor W (eds.) *Proceedings of the 11th International Workshop on Neural-Symbolic Learning and Reasoning (NeSy'16) co-located with the Joint Multi-Conference on Human-Level Artificial Intelligence (HLAI 2016), New York City, NY, USA, July 16-17, 2016*, *CEUR Workshop Proceedings*, volume 1768. CEUR-WS.org. URL https://ceur-ws.org/Vol-1768/NESY16_paper3.pdf.

Sharma S, Zhang Y, Aliaga JMR, Bouneffouf D, Muthusamy V and Varshney KR (2020) Data augmentation for discrimination prevention and bias disambiguation. In: Markham AN, Powles J, Walsh T and Washington AL (eds.) *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*. ACM, pp. 358–364. DOI:10.1145/3375627.3375865. URL https://doi.org/10.1145/3375627.3375865.

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap TP, Leach M, Kavukcuoglu K, Graepel T and Hassabis D (2016) Mastering the game of go with deep neural networks and tree search. *Nat.* 529(7587): 484–489. DOI:10.1038/NATURE16961. URL https://doi.org/10.1038/nature16961.

Simson J, Fabris A and Kern C (2024) Lazy data practices harm fairness research. In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024, Rio de Janeiro, Brazil, June 3-6, 2024*. ACM, pp. 642–659. DOI:10.1145/3630106.3658931. URL https://doi.org/10.1145/3630106.3658931.

Stanovich KE and West RF (2000) Advancing the rationality debate. *Behavioral and Brain Sciences* 23(5): 701–717. DOI:10.1017/S0140525X00623439.

Suriyakumar VM, Ghassemi M and Ustun B (2023) When personalization harms performance: Reconsidering the use of group attributes in prediction. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S and Scarlett J (eds.) *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, *Proceedings of Machine Learning Research*, volume 202. PMLR, pp. 33209–33228. URL https://proceedings.mlr.press/v202/suriyakumar23a.html.

Tavakol M (2020) Fair classification with counterfactual learning. In: Huang JX, Chang Y, Cheng X, Kamps J, Murdock V, Wen J and Liu Y (eds.) *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*. ACM, pp. 2073–2076. DOI:10.1145/3397271.3401291. URL https://doi.org/10.1145/3397271.3401291.

Valdivia A, Sánchez-Monedero J and Casillas J (2021) How fair can we go in machine learning? assessing the boundaries of accuracy and fairness. *Int. J. Intell. Syst.* 36(4):

1619–1643. DOI:10.1002/INT.22354. URL https://doi.org/10.1002/int.22354.

Wagner B and d'Avila Garcez AS (2021) Neural-symbolic integration for fairness in AI. In: Martin A, Hinkelmann K, Fill H, Gerber A, Lenat D, Stolle R and van Harmelen F (eds.) *Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021), Stanford University, Palo Alto, California, USA, March 22-24, 2021*, *CEUR Workshop Proceedings*, volume 2846. CEUR-WS.org. URL https://ceur-ws.org/Vol-2846/paper5.pdf.

Wan M, Zha D, Liu N and Zou N (2023) In-processing modeling techniques for machine learning fairness: A survey. *ACM Trans. Knowl. Discov. Data* 17(3): 35:1–35:27. DOI:10.1145/3551390. URL https://doi.org/10.1145/3551390.

Wan Z, Liu C, Yang H, Raj R, Li C, You H, Fu Y, Wan C, Samajdar A, Lin YC, Krishna T and Raychowdhury A (2024) Towards cognitive AI systems: Workload and characterization of neuro-symbolic AI. In: *IEEE International Symposium on Performance Analysis of Systems and Software, ISPASS 2024, Indianapolis, IN, USA, May 5-7, 2024*. IEEE, pp. 268–279. DOI:10.1109/ISPASS61541.2024.00033. URL https://doi.org/10.1109/ISPASS61541.2024.00033.

Wang J, Wang XE and Liu Y (2022) Understanding instance-level impact of fairness constraints. In: Chaudhuri K, Jegelka S, Song L, Szepesvári C, Niu G and Sabato S (eds.) *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, *Proceedings of Machine Learning Research*, volume 162. PMLR, pp. 23114–23130. URL https://proceedings.mlr.press/v162/wang22ac.html.

Wang X and Huang H (2019) Approaching machine learning fairness through adversarial network. *CoRR* abs/1909.03013. URL http://arxiv.org/abs/1909.03013.

Wirtz BW, Weyerer JC and Geyer C (2019) Artificial intelligence and the public sector—applications and challenges. *International Journal of Public Administration* 42(7): 596–615. DOI:10.1080/01900692.2018.1498103. URL https://doi.org/10.1080/01900692.2018.1498103.

Wu C, Wu F, Qi T and Huang Y (2022) Semi-fairvae: Semi-supervised fair representation learning with adversarial variational autoencoder. *CoRR* abs/2204.00536. DOI:10.48550/ARXIV.2204.00536. URL https://doi.org/10.48550/arXiv.2204.00536.

Xu D, Wu Y, Yuan S, Zhang L and Wu X (2019a) Achieving causal fairness through generative adversarial networks. In: Kraus S (ed.) *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. ijcai.org, pp. 1452–1458. DOI:10.24963/IJCAI.2019/201. URL https://doi.org/10.24963/ijcai.2019/201.

Xu D, Yuan S, Zhang L and Wu X (2018) Fairgan: Fairness-aware generative adversarial networks. In: Abe N, Liu H, Pu C, Hu X, Ahmed NK, Qiao M, Song Y, Kossmann D, Liu B, Lee K, Tang J, He J and Saltz JS (eds.) *IEEE International Conference on Big Data (IEEE BigData 2018), Seattle, WA, USA, December 10-13, 2018*. IEEE, pp. 570–575. DOI:10.1109/BIGDATA.2018.8622525. URL https://doi.org/10.1109/BigData.2018.8622525.

Xu D, Yuan S, Zhang L and Wu X (2019b) Fairgan$^+$: Achieving fair data generation and classification through generative adversarial nets. In: Baru CK, Huan J, Khan

L, Hu X, Ak R, Tian Y, Barga RS, Zaniolo C, Lee K and Ye YF (eds.) *2019 IEEE International Conference on Big Data (IEEE BigData), Los Angeles, CA, USA, December 9-12, 2019*. IEEE, pp. 1401–1406. DOI:10.1109/BIGDATA47090.2019.9006322. URL https://doi.org/10.1109/BigData47090.2019.9006322.

Zehlike M, Hacker P and Wiedemann E (2020) Matching code and law: achieving algorithmic fairness with optimal transport. *Data Min. Knowl. Discov.* 34(1): 163–200. DOI:10.1007/S10618-019-00658-8. URL https://doi.org/10.1007/s10618-019-00658-8.

Zemel RS, Wu Y, Swersky K, Pitassi T and Dwork C (2013) Learning fair representations. In: *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, *JMLR Workshop and Conference Proceedings*, volume 28. JMLR.org, pp. 325–333. URL http://proceedings.mlr.press/v28/zemel13.html.

Zhang L, Wu Y and Wu X (2017) A causal framework for discovering and removing direct and indirect discrimination. In: Sierra C (ed.) *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. ijcai.org, pp. 3929–3935. DOI:10.24963/IJCAI.2017/549. URL https://doi.org/10.24963/ijcai.2017/549.

Zhang W and Weiss JC (2023) Fair decision-making under uncertainty. *CoRR* abs/2301.12364. DOI:10.48550/ARXIV.2301.12364. URL https://doi.org/10.48550/arXiv.2301.12364.

Zliobaite I, Kamiran F and Calders T (2011) Handling conditional discrimination. In: Cook DJ, Pei J, Wang W, Zaïane OR and Wu X (eds.) *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011*. IEEE Computer Society, pp. 992–1001. DOI:10.1109/ICDM.2011.72. URL https://doi.org/10.1109/ICDM.2011.72.