Fairness-Aware Knowledge Distillation via Symbolic Knowledge Extractors

Federico Sabbatini[1,2,*], Roberta Calegari[1]

[1]Department of Computer Science and Engineering, University of Bologna, Viale del Risorgimento 2, 40136, Bologna, Italy
[2]National Institute for Nuclear Physics, Section in Florence, Via Bruno Rossi 1, 50019, Sesto Fiorentino, Italy

Abstract

As machine learning systems are increasingly deployed in sensitive domains, ensuring both fairness and interpretability has become a critical requirement. Symbolic knowledge-extraction techniques offer human-readable surrogate models for black-box predictors but often lack mechanisms to guarantee fairness. This paper introduces fairness-aware extensions to several knowledge-extraction algorithms within the PSyKE platform. By excluding user-specified sensitive attributes during the extraction process, the resulting symbolic models maintain interpretability while promoting fairness—even when derived from biased data or predictors. Experimental results on the Adult data set show that these extensions eliminate counterfactual unfairness and improve disparate impact, without sacrificing predictive accuracy. This work advances the state of the art in trustworthy and transparent artificial intelligence by integrating fairness directly into the symbolic model extraction pipeline.

Keywords

Algorithmic fairness, XAI, Symbolic knowledge, PSyKE.

1. Introduction

The increasing deployment of machine learning (ML) systems in high-stakes domains – e.g., healthcare, finance, and criminal justice – has raised significant concerns about fairness and interpretability. Ensuring that ML models provide equitable outcomes for individuals across different demographic groups is essential to prevent discrimination and uphold ethical standards. At the same time, interpretability is critical for transparency, user trust, and regulatory compliance, allowing stakeholders to understand, audit, and contest automated decisions. Symbolic and inherently interpretable models have gained renewed attention due to their transparency and explanatory capabilities. These models encode decision logic in human-readable forms – such as rules, trees, or propositional logic – enabling domain experts and affected users to understand the rationale behind automated decisions. However, symbolic models often rely on training data or black-box predictors that may embed structural or statistical biases related to sensitive attributes such as race, gender, or age. As a result, the symbolic models extracted from such sources risk perpetuating or even amplifying unfairness, unless explicitly designed to counteract it.

The ML community has recently proposed a wide range of fairness-enhancing interventions, including pre-processing methods that modify the data to remove bias, in-processing methods to incorporate fairness constraints during model training, and post-processing methods that adjust predictions to mitigate unfair outcomes [1, 2]. Explainable artificial intelligence (XAI) has also been explored as a tool for fairness auditing and mitigation. *Post-hoc* explanation methods, such as LIME [3], SHAP [4], and counterfactual reasoning [5], are often used to reveal the influence of sensitive features and to identify disparate treatment. However, recent work has highlighted the limitations of these approaches, including vulnerability to manipulation [6], inconsistency [7], and lack of robustness in fairness-critical applications [8]. While these techniques have significantly advanced the field, most fairness-aware solutions treat interpretability and fairness as separate concerns—either applying fairness constraints during model training or attempting to diagnose unfairness *post hoc* via explanations. In contrast,

2nd European Workshop in Trustworthy (TRUST-AI 2026), August 15–17, 2026, Bremen, Germany
*Corresponding author.
f.sabbatini@unibo.it (F. Sabbatini); roberta.calegari@unibo.it (R. Calegari)
0000-0002-0532-6777 (F. Sabbatini); 0000-0003-3794-2942 (R. Calegari)

symbolic knowledge extraction (SKE) offers a unique opportunity to unify these objectives by producing inherently interpretable surrogates that are also fair by design [9, 10]. However, existing SKE methods lack mechanisms to enforce fairness during extraction, limiting their reliability in sensitive domains.

This paper addresses the challenge of producing fair and interpretable models through the extraction of symbolic knowledge from opaque, potentially biased predictors, without altering the original data set or modifying the underlying predictive models. We leverage the PSyKE platform [11, 12, 13] to introduce fairness-aware extensions to several SKE algorithms. By allowing users to specify protected or sensitive attributes to be excluded during the extraction process, we ensure that the resulting symbolic models do not rely on protected features, thereby promoting fairness in the extracted knowledge. The result is a surrogate model that is faithful to the original predictions, interpretable by design, and fair—even when derived from biased data or black-box predictors.

Our contribution includes the adaptation of multiple extraction algorithms – spanning decision trees, hypercubic partitioning, and interpretable clustering – to incorporate fairness constraints at extraction time. These enhancements are seamlessly integrated into the PSyKE platform, which provides a unified and extensible interface for generating fair symbolic surrogates from arbitrary black-box models. To the best of our knowledge, this work represents the first general-purpose extension of SKE techniques to support fairness and bias mitigation. By adopting an in-processing strategy, our approach offers a novel avenue for fairness in explainable ML, particularly in scenarios where modifying the training data or black-box predictions is impractical or prohibited. This work contributes to the broader goal of developing trustworthy, transparent, and equitable intelligent systems that align with ethical principles and regulatory requirements.

2. Related Works

In the broader fairness literature, techniques to mitigate bias in systems based on artificial intelligence are typically classified into three categories: pre-processing methods, which modify the data before training; in-processing methods, which alter the training procedure to enforce fairness constraints; and post-processing methods, which adjust the model outputs after learning [14]. While XAI is often used as a post-processing tool to diagnose potential discrimination, our approach instead falls into the in-processing category, enforcing fairness during the generation of symbolic surrogates.

XAI has long been regarded as a potential remedy for the opacity of ML models and the resulting challenges around fairness, accountability, and trust. Yet, recent literature calls for a more critical appraisal of this assumption [7], arguing for a clearer distinction between epistemic uses of XAI, such as enabling audits or bias detection, and substantial uses, which involve modifying or constraining the model to promote fairness. This distinction is key for our work: whereas prior approaches often deploy XAI as a tool to inspect already-trained models, our approach embeds fairness directly into SKE.

In parallel, important threads in the fairness literature focused on fairness-aware knowledge distillation [8, 15]. [15] proposes a knowledge distillation framework using soft labels from an overfitted teacher model to guide the training of a fairer student model without explicit access to demographic information. The method effectively operationalises fairness-through-reweighing and achieves empirical improvements in group fairness metrics across several benchmark data sets. However, the resulting models are still black-box learners, and the fairness they achieve is implicit. In contrast, our contribution brings fairness into the symbolic domain by enforcing fairness constraints during the extraction process itself. This ensures that the resulting interpretable models not only reproduce the behaviour of the black-box teacher but also do so without relying on protected attributes or their proxies.

The interplay between explainability and fairness is also examined in studies that treat XAI as a lens for auditing existing models. [8] provides a comprehensive investigation of whether *post-hoc* explanation methods – such as LIME, SHAP, and DiCE – can reliably detect and characterise discriminatory patterns in model behaviour. The results in [8] confirm that explanation methods can reveal disparities in feature attributions across groups and expose indirect discrimination even when protected attributes are removed. However, they also emphasise the limitations of XAI in this role: explanations can

be fragile, easily manipulated, and context-dependent. Proxies for sensitive attributes frequently emerge when those attributes are omitted, suggesting that removing sensitive features is insufficient for guaranteeing fairness. These insights underscore the limitations of *post-hoc* strategies and motivate our in-processing approach, where fairness is not merely assessed but structurally enforced at the point of model generation.

Additional concerns about the limits of XAI for fairness have been raised in [6], where it is demonstrated how LIME and SHAP explanations can be adversarially manipulated to "fairwash" discriminatory models. Similarly, in [16] it is argued that *post-hoc* explanations of black-box models are often unfaithful to the underlying logic and that intrinsically interpretable models should be preferred in high-stakes contexts. Our work resonates with this argument but adds a new dimension: rather than designing transparent models from scratch, we extract fair symbolic approximations from existing black-box models, enabling interpretability and fairness to coexist without retraining or data alteration.

Finally, fairness-aware knowledge distillation has recently been explored in ranking and recommendation systems. [17] presents a fairness ranking distillation method that integrates exposure-based constraints into the distillation loss to ensure fair visibility of items. While the approach is notable for reducing model size and improving exposure fairness, it is tailored to recommender systems and does not address interpretability. Our work extends the distillation paradigm to the domain of symbolic models, uniting the principles of fairness, fidelity, and interpretability in a common extraction pipeline.

Across these lines of inquiry, a recurring limitation is the absence of methods that ensure fairness during the generation of interpretable models. While existing literature provides tools for auditing, mitigating, or approximating fairness, few works propose fairness-aware algorithms for SKE. Our contribution fills this gap by introducing an integrated framework that allows users to specify fairness constraints – namely, the exclusion of sensitive and protected attributes – and obtain symbolic surrogates that are by design faithful to the black-box model and free from sensitive feature dependencies.

3. Fairness-Aware Symbolic Knowledge Extraction

This paper presents an extension of SKE algorithms by introducing fairness-aware capabilities. While the underlying extraction methods are pre-existing, our key contribution lies in their systematic re-engineering to explicitly incorporate bias mitigation constraints within the knowledge distillation process. Introducing a mechanism to exclude user-defined protected or sensitive attributes from the knowledge extraction process cannot be achieved uniformly. Our goal is to prevent the encoding of bias present in the training data or introduced during model learning by means of an in-processing strategy. This requires the implementation of *ad-hoc* modifications tailored to each extraction algorithm, depending on its specific logic and structure.

To maximise the applicability of our methodology, we focus on several pedagogical extractors – i.e., those that operate independently of the internal structure of the underlying black-box model – belonging to different methodological families. These include CART [18] (decision tree-based), ITER[19] and GridEx [20] (based on hypercubic partitioning of the input domain; [21]), and CReEPy [22] (interpretable clustering). To facilitate the integration and evaluation of these heterogeneous SKE techniques, we extend the PSyKE platform, positioning it as a unified hub for fairness-aware SKE. This design choice ensures consistency, reusability, and ease of adoption across diverse application domains and algorithmic backends. The integration of fairness constraints within the extraction process preserves full compatibility with the PSyKE architecture and its high-level abstractions, offering an extensible and user-friendly interface that enables developers and practitioners to generate fair, interpretable models from opaque predictors with minimal integration effort.

3.1. Integrating Fairness Constraints into SKE

CART. Although not originally conceived as an SKE, CART [18] produces decision trees in which each path from the root to a leaf corresponds to a complete and self-contained classification or regression rule. This structure results in an inherently interpretable model that can be effectively employed

to approximate the decision boundaries of otherwise opaque systems. The core algorithm follows a recursive, top-down construction process, starting from the root node, which is associated with the entire data set. At each internal node, the optimal split is selected based on a predefined impurity metric and the tree is expanded accordingly. This splitting continues until one or more stopping criteria are met, such as a maximum depth, a minimum number of samples per leaf, or a minimum impurity decrease. To mitigate overfitting and enhance generalisation, pruning strategies can be applied *post hoc* to remove superfluous branches, simplifying the model while preserving predictive accuracy. CART supports both classification and regression and accommodates a broad range of inputs, including continuous and categorical features. CART, in its standard form, does not include any bias mitigation mechanism: all input features are eligible for selection during tree growth, and no fairness-aware terms are added to the impurity evaluation to penalise biased splits. To address fairness concerns in decision tree induction, our extension of CART explicitly excludes sensitive or protected attributes from consideration during the construction process. At each split, the set of candidate features is dynamically filtered to exclude those identified by the user as sensitive. The algorithm then proceeds as in standard CART, evaluating impurity reductions over the admissible subset of features. This constraint ensures that no decision rule in the resulting tree directly depends on protected attributes, thereby mitigating the risk of disparate treatment and enhancing the interpretability of the model from a fairness perspective. The method remains compatible with standard pruning strategies and supports both classification and regression.

CReEPy. CReEPy [22] is an SKE algorithm that leverages explainable clustering to derive interpretable rule-based models from arbitrary black-box classifiers or regressors. Clustering algorithms are first applied, then the algorithm recursively constructs a hierarchical partitioning of the input feature space using axis-aligned hypercubes approximating the clusters. This structure is ultimately translated into a propositional rule tree that serves as a human-readable representation of the black-box decision logic. The hypercubic subregions identified by CReEPy are converted into propositional rules, where each rule antecedent corresponds to the conjunction of conditions defining the associated cube. A relevance-based filtering step follows, in which input variables deemed insufficiently informative are pruned from the rule preconditions to enhance conciseness and readability. To promote fairness in CReEPy, our extension explicitly excludes sensitive or protected features during the clustering phase. As a result, the identification and approximation of clusters are based solely on admissible features, and the derived rules are constructed exclusively over the non-sensitive subspace.

ITER. ITER [19] is designed to extract human-interpretable rules from opaque regression models operating on continuous input domains. The method constructs axis-aligned hypercubes in the input space, which are subsequently converted into propositional rule lists approximating the behaviour of the target model. The algorithm builds each hypercube through a greedy, similarity-driven expansion process. Starting from a point in the input space, a cube is iteratively expanded along the dimension that maximises similarity with neighbouring points, effectively growing the region while preserving output homogeneity. This leads to non-overlapping partitions in which each hypercube corresponds to a group of instances with similar predicted values. Each hypercube is then converted into a rule mapping a conjunction of feature conditions to a constant output value [23]. To integrate fairness constraints into ITER, our extension forbids expansion along sensitive or protected dimensions. At each step, candidate directions are filtered to retain only admissible features, and similarity is computed in the non-sensitive subspace. This ensures that the resulting rule-based model avoids reliance on protected attributes and mitigates the risk of encoded bias.

GridEx. GridEx [20] is a recursive SKE algorithm that partitions the input space into axis-aligned hypercubes. It follows a top-down strategy driven by both feature relevance and the local fidelity of the black-box model. Features deemed important receive finer grid resolutions, and regions where the model already performs well are left unpartitioned to avoid unnecessary complexity. Each resulting cube is associated with an output value and translated into a propositional rule. To ensure fairness, our extension of GridEx prohibits the use of sensitive or protected features as partitioning axes. The splitting mechanism is constrained to operate solely over the user-defined admissible feature set. Consequently, the generated rule set reflects decision logic derived exclusively from non-sensitive attributes, thereby improving its fairness while retaining the interpretability and fidelity guarantees of the original method.

Listing 1: Rules extracted with CART for the Adult data set.

```
income(Education, ..., Age, Sex, '<=50K') :- Education =< 12.5.
income(Education, ..., Age, Sex, '>50K')  :- Education > 12.5, Sex = 'M'.
income(Education, ..., Age, Sex, '<=50K') :- Education > 12.5, Sex = 'F'.
```

3.2. Preferring In-Processing over Post-Processing

It is emphasised that the proposed extensions are not the only way to obtain knowledge bases both human-interpretable and fair. For instance, post-processing routines aimed at removing knowledge items deemed biased may also be applied. However, such strategies often introduce significant drawbacks, including reduced coverage of the input space and degraded predictive performance. Specifically, if knowledge items are disjoint and no default rule is defined for fallback prediction, the removal of a rule may leave a portion of the input domain uncovered. Conversely, if a default rule is present, all uncovered instances will be assigned to it, potentially resulting in poor predictive accuracy due to oversimplification. Another post-processing strategy involves removing subcomponents of the knowledge items—such as individual preconditions in *if-then-else* rules. However, this may produce unintended consequences, including duplicate rules with identical antecedents but conflicting consequents, thereby yielding inconsistent or contradictory predictions. An illustrative example is provided in Listing 1, where a knowledge base is extracted from an ML model trained on the Adult data set (https://archive.ics.uci.edu/dataset/2/adult). In this case, the second and third rules differ only in the use of the `sex` attribute, which is considered sensitive. Pruning both rules would yield a knowledge base that provides only negative outcomes based on the `education` feature. Pruning only the sensitive condition (i.e., the `sex` attribute) from both rules would result in rules with identical preconditions but contradictory predictions, thus violating consistency. In light of these considerations, we focus on in-processing strategies despite their greater complexity in terms of design, implementation, and integration within existing fairness-unaware algorithms.

4. Experiments

The effectiveness of the proposed fairness-oriented extensions is demonstrated through an experimental evaluation conducted on the Adult data set. This benchmark consists of 14 input features, both numerical and categorical, used to predict whether an individual's annual income exceeds $50,000. The target variable is thus binary, distinguishing between high-income and low-income classes. The disparate impact index (DI; [24]) computed on the raw Adult data set was found to be 0.36, which, according to established fairness thresholds, reflects a substantial disparity in treatment between male and female individuals (see Table 1). We recall that this metric quantifies the degree of unequal treatment between two demographic groups by comparing the proportion of individuals receiving favourable outcomes in each group. Specifically, it yields a numerical indication of potential bias or discrimination, with values closer to 1 suggesting parity, and values significantly below 1 indicating possible unfairness. The DI can be evaluated at various stages of the ML pipeline: on the raw data set, to assess pre-existing bias; on the predictions of a trained black-box model, to determine whether the learning process introduces, mitigates, or amplifies unfairness; and on the symbolic knowledge extracted from the model, to verify whether the fairness level is preserved in the interpretable surrogate.

Following the initial assessment, 4 opaque ML models were trained to evaluate how the original bias was affected during the learning process. A training/test split of 75%/25% was adopted. The models considered include: a random forest (RF) classifier composed of 25 unbounded trees; a K-nearest neighbours (K-NN) classifier with $K = 150$; a Gaussian naïve Bayes (GNB) model; and a feed-forward fully-connected artificial neural network (ANN) with two hidden layers of 75 and 25 neurons, respectively, employing ReLU activations. The DI was computed for the predictions of all these models. The resulting values, shown in Table 1, ranged from 0.07 to 0.45, indicating a persistent and in some cases exacerbated level of unfairness. These results suggest that the discriminatory patterns

Table 1
Results. The DI is computed on the raw data, on the predictions of ML models, and on symbolic knowledge bases obtained via standard and fairness-aware extraction techniques. CFF is measured as the proportion of individuals whose predicted outcome changes when the sex attribute is flipped. DI is measured as the proportion of female individuals receiving favourable outcomes over the proportion of male individuals receiving the same outcome

	Accuracy		CFF		DI	
	Orig.	Fair	Orig.	Fair	Orig.	Fair
Data set	-	-	-	-	0.360	-
Training set	-	-	-	-	0.360	-
Test set	-	-	-	-	0.358	-
RF	0.822	-	0.372	-	0.203	-
+ CART	0.811	0.788	**0.000**	**0.000**	0.406	**0.768**
+ CReEPy	0.791	0.795	0.527	**0.000**	0.011	0.572
+ ITER	0.783	0.777	0.503	**0.000**	0.000	0.586
+ GridEx	0.770	0.771	0.508	**0.000**	0.006	0.576
K-NN	0.836	-	0.461	-	0.066	-
+ CART	0.811	0.789	0.573	**0.000**	0.030	**0.705**
+ CReEPy	0.773	0.774	0.482	**0.000**	0.055	0.536
+ ITER	0.772	0.772	**0.000**	**0.000**	0.586	0.595
+ GridEx	0.780	0.783	0.539	**0.000**	0.037	0.624
GNB	0.800	-	0.450	-	0.448	-
+ CART	0.803	0.803	**0.000**	**0.000**	0.511	0.521
+ CReEPy	0.793	0.782	0.524	**0.000**	0.016	0.593
+ ITER	0.774	0.771	0.504	**0.000**	0.000	0.570
+ GridEx	0.779	0.778	0.542	**0.000**	0.000	**0.597**
ANN	0.845	-	0.323	-	0.174	-
+ CART	0.812	0.793	**0.000**	**0.000**	0.645	**0.675**
+ CReEPy	0.769	0.794	0.515	**0.000**	0.000	0.566
+ ITER	0.770	0.772	0.507	**0.000**	0.000	0.586
+ GridEx	0.781	0.781	**0.000**	**0.000**	0.596	0.596

present in the original data were not only preserved but, in several cases, further amplified by the learned decision functions. This underscores the importance of integrating fairness-aware mechanisms during model training or within the subsequent knowledge extraction phase, with the ultimate goal of obtaining a fair and interpretable predictive model.

In addition to the DI analysis, counterfactual fairness (CFF; [25]) was evaluated for each of the trained black-box models. This metric estimates the proportion of individuals whose predicted outcome changes when a sensitive attribute – in this case, sex – is altered, while all other input features are held constant. Such counterfactual analysis provides a direct measure of predictive bias, revealing the extent to which a model's decisions rely on protected characteristics. In this work, CFF is computed as the ratio of individuals whose prediction changes when their sex attribute is flipped, relative to the total number of individuals. For instance, a CFF score of 0.15 indicates that 15% of individuals receive a different outcome solely due to their sex attribute. The results show that a non-negligible proportion of instances, ranging from 32% to 46%, exhibit a change in predicted outcome when the sex attribute is modified. This confirms the presence of counterfactual unfairness across all models. These findings further reinforce the conclusion that the learned decision functions are sensitive to protected attributes and underscore the need for fairness-aware *post-hoc* interpretability methods capable of exposing and mitigating biases.

Listing 2: Rules extracted with standard CART applied to the K-NN.

```
income(Education, ..., Gain, Sex, '>50K')  :- Education > 12.5, Sex = 'M'.
income(Education, ..., Gain, Sex, '<=50K') :- Education > 12.5, Sex = 'F'.
income(Education, ..., Gain, Sex, '<=50K') :- Gain =< 7073.50.
income(Education, ..., Gain, Sex, '>50K').
```

Listing 3: Rules extracted with fair-oriented CART applied to the K-NN.

```
income(Education, ..., Age, '<=50K') :- Education =< 11.0.
income(Education, ..., Age, '<=50K') :- Age =< 34.0.
income(Education, ..., Age, '>50K').
```

4.1. Interpretable and Fair Symbolic Knowledge Extraction

To investigate the impact of SKE on fairness, both the standard and fairness-aware versions of the extractors described in the previous section – namely, CART, CReEPy, ITER, and GridEx – were applied to each of the considered black-box models. This resulted in a total of 32 interpretable models, each approximating one of the opaque classifiers through symbolic knowledge. The majority of standard extractors yielded models affected by approximately 50% counterfactual unfairness, meaning that one out of two individuals received a prediction dependent on the `sex` attribute. In contrast, the fairness-aware versions of the extractors consistently achieved a counterfactual fairness score of zero, thereby ensuring counterfactually fair outcomes.

The DI values measured for the symbolic knowledge bases never exceeded the 0.8 threshold commonly used in the United States to distinguish between fair and unfair models [24]; nonetheless, notable improvements were observed. These results point to the presence of hidden correlations between protected attributes and the remaining admissible features, which may continue to influence predictions indirectly. Notably, the DI values for standard extractors are typically very low (often close to 0), with rare exceptions reaching between 0.5 and 0.6. Conversely, the fairness-oriented variants consistently yielded higher DI scores, in some cases peaking at 0.77. These findings indicate that the proposed fairness extensions are already effective at improving fairness in symbolic models, although they are not yet sufficient for full mitigation of all bias sources.

As an example, Listings 2 and 3 presents knowledge bases extracted using the standard and fairness-aware versions of CART applied to the opaque K-NN classifier. Listing 2 shows a biased rule set, in which positive outcomes are assigned to male individuals with high education levels, while female individuals with the same educational background receive negative outcomes. Additionally, individuals with low income are always classified negatively, whereas the remaining individuals are given positive outcomes. The first two rules clearly reveal a direct dependence on the `sex` attribute, indicating unfair behaviour. By contrast, Listing 3 shows the knowledge extracted by the fairness-aware variant of CART on the same black-box model. Here, negative outcomes are assigned to individuals with low education levels or young individuals, regardless of their education. Positive predictions are reserved for the remaining cases. If additional attributes such as `age` are also deemed sensitive, they can be excluded alongside `sex` and `race`, resulting in a different knowledge base that adheres to revised fairness criteria.

Further insight can be gained by examining the heatmaps shown in Figure 1, which visualise rule coverage across sensitive groups for the knowledge bases in Listings 2 and 3. The heatmap corresponding to the standard CART model reveals clear unfairness in the first two rules, as indicated by noticeable differences in background colour between male and female groups for the same rule—where fairness would require similar colouring. Additionally, sex-based disparities are also visible in the third and fourth rules. The fairness-aware CART model, while significantly more equitable, still exhibits residual bias. For example, the second rule disproportionately assigns negative outcomes to female individuals, with a 7% higher probability compared to males under the same rule conditions. Conversely, the third rule favours male individuals, who are 4% more likely to receive a positive outcome than their female counterparts. These observations confirm that even when sensitive features are excluded from the

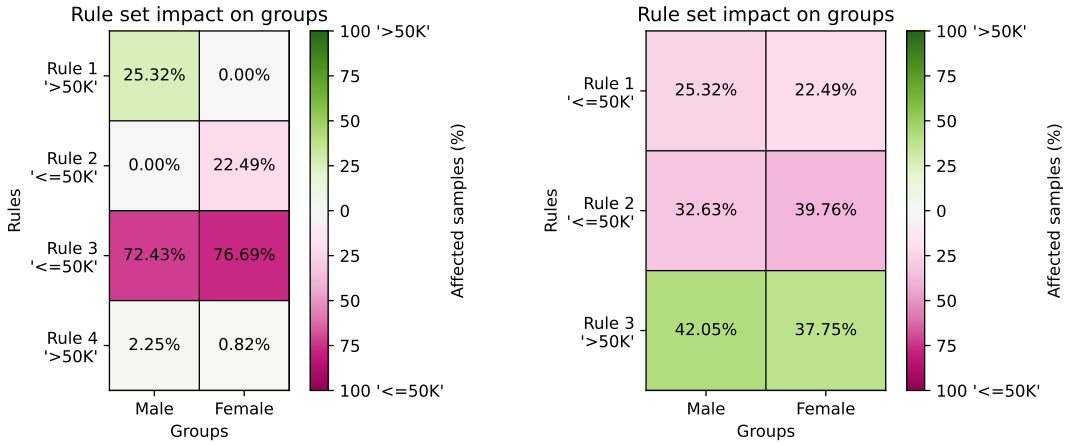

Figure 1: Coverage across sensitive groups for the knowledge bases reported in Listing 2 and Listing 3.

Table 2
Consistency experiments. Same as Table 1 for the Loan data set.

	Accuracy		CFF		DI	
	Orig.	Fair	Orig.	Fair	Orig.	Fair
Data set	-	-	-	-	0.890	-
Training set	-	-	-	-	0.891	-
Test set	-	-	-	-	0.873	-
RF	0.788	-	0.000	-	0.989	-
+ CART	0.783	0.783	0.000	0.000	0.989	0.989
+ CReEPy	0.750	0.750	0.000	0.000	1.000	1.000
+ ITER	0.690	0.690	0.000	0.000	1.000	1.000
+ GridEx	0.663	0.663	0.000	0.000	1.000	1.000
ANN	0.785	-	0.000	-	1.000	-
+ CART	0.786	0.786	0.000	0.000	1.000	1.000
+ CReEPy	0.751	0.751	0.000	0.000	1.000	1.000
+ ITER	0.793	0.793	0.000	0.000	1.000	1.000
+ GridEx	0.660	0.660	0.000	0.000	1.000	1.000

symbolic model, indirect forms of unfairness may still emerge due to latent correlations in the data.

4.2. Consistency Experiments

To verify that the fairness-aware extensions do not introduce unnecessary changes in unbiased settings, we conducted control experiments using the Loan data set (https://www.kaggle.com/datasets/burak3ergun/loan-data-set), which does not exhibit gender- or ethnicity-based unfairness. This allowed us to assess whether the modified SKE algorithms behave consistently with their original counterparts when no mitigation is required. Two black-box ML models were used: an RF with 35 trees and a maximum depth of 4, and an ANN with two hidden layers of 70 and 30 neurons, respectively, using ReLU activation functions. A training/test split of 75%/25% was adopted, consistent with the previous case study. The data set exhibits a DI score of 0.89, as shown in Table 2. Both black-box models achieved an accuracy of approximately 0.79, with DI values close to 1 and CFF scores of 0—indicating unbiased predictions aligned with the fairness of the training data. Comparable fairness and accuracy metrics were obtained for the symbolic knowledge bases extracted via CART, GridEx, ITER, and CReEPy. As sensitive attributes were not excluded during the extraction process in this case – since no unfairness was present – no differences emerged between the original SKE algorithms and their fairness-aware

counterparts. These results, reported in Table 2, confirm that the proposed extensions do not introduce unnecessary changes when applied in bias-free settings.

4.3. Discussion and Future Developments

Fairness is a multifaceted and nuanced concept, which is reflected in the existence of multiple metrics designed to capture different dimensions of unfairness. These metrics may yield conflicting assessments, making it particularly challenging to devise bias mitigation strategies that are universally effective across all fairness notions. This is evident in the results reported in Table 1, where certain predictors exhibit low DI despite being 100% counterfactually fair.

The classification accuracy of the fairness-aware extractors is consistently comparable to that of their standard counterparts, suggesting that the proposed extensions improve fairness without sacrificing predictive performance. However, these results should not be interpreted as providing a complete solution to algorithmic unfairness. In particular, the proposed approach is insufficient for fully addressing proxy-based or indirect discrimination, since eliminating explicit dependence on sensitive attributes does not guarantee the removal of all unfair decision patterns.

This limitation is especially relevant in the context of pedagogical SKE. Since the proposed fairness-aware extensions operate by extracting symbolic predictors that approximate the behaviour of a potentially biased black-box while excluding decisions directly driven by sensitive or protected variables, the extracted models are not intended to be exact replicas of the original predictor. As a consequence, fidelity with respect to the source black-box may decrease whenever unfair decisions are intentionally altered or flipped in favour of fairer outcomes. The resulting symbolic representations therefore aim to preserve predictive behaviour while reducing counterfactual bias, rather than maximising agreement with the original model. This also explains why CFF and DI may not necessarily evolve consistently across extractors: removing explicit sensitivity to protected attributes does not prevent residual unfairness induced by indirect decision mechanisms.

The proposed approach does not limit the impact of proxy or hidden biases. Non-sensitive attributes may still encode information strongly correlated with protected characteristics and therefore act as indirect channels through which biased decision patterns are reproduced. Since the proposed extensions primarily intervene on explicit dependence structures, they may not always possess sufficient expressive capacity to identify and neutralise such latent associations. Consequently, even in the presence of complete CFF, DI may remain non-negligible. Addressing this issue would likely require fairness-aware training objectives capable of detecting and penalising indirect discrimination mechanisms.

A promising direction for future development involves the integration of user-defined, fairness-aware penalty terms within the training procedures of SKE algorithms. For example, such penalties could influence the selection of splitting criteria in CART, the expansion strategy in ITER, or the decision to further partition hypercubes in GridEx and CReEPy. This would enable more fine-grained control over the fairness-accuracy trade-off and could potentially mitigate indirect bias arising from proxy variables. Future studies will also focus on broadening the experimental evaluation to other benchmark data sets.

5. Conclusions

This paper presents novel methods for generating fairness-aware symbolic models by modifying existing knowledge extraction algorithms within the PSyKE platform. By excluding sensitive or protected attributes during the extraction process, the resulting symbolic models remain both interpretable and fair—even when derived from biased data or black-box predictors. Aligned with the in-processing paradigm, the proposed methods do not require modifications to the original data set or *post-hoc* adjustments to model predictions. Their integration into PSyKE provides a unified interface that enables practitioners to generate, compare, and adapt fair symbolic models efficiently. Future work will focus on extending fairness constraints to support more complex symbolic representations, to broaden the applicability of the proposed approach and improve its flexibility and overall performance.

Declaration on Generative AI

During the preparation of this work, the authors used Generative AI tools for grammar and spell checking. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Barocas, M. Hardt, A. Narayanan, Fairness and machine learning: Limitations and opportunities, MIT press, 2023.

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM computing surveys (CSUR) 54 (2021) 1–35.

[3] M. T. Ribeiro, S. Singh, C. Guestrin, Model-agnostic interpretability of machine learning, arXiv preprint arXiv:1606.05386 (2016).

[4] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[5] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech. 31 (2017) 841.

[6] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, Fooling lime and shap: Adversarial attacks on post hoc explanation methods, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 180–186.

[7] L. Deck, J. Schoeffer, M. De-Arteaga, N. Kühl, A critical survey on fairness benefits of explainable ai, in: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024, pp. 1579–1595.

[8] V. Papanikou, D. P. Karidi, E. Pitoura, E. Panagiotou, E. Ntoutsi, Explanations as bias detectors: A critical study of local post-hoc xai methods for fairness exploration, arXiv preprint arXiv:2505.00802 (2025).

[9] G. Ciatto, F. Sabbatini, A. Agiollo, M. Magnini, A. Omicini, Symbolic knowledge extraction and injection with sub-symbolic predictors: A systematic literature review, ACM Computing Surveys 56 (2024) 161:1–161:35. doi:10.1145/3645103.

[10] F. Sabbatini, Four decades of symbolic knowledge extraction from sub-symbolic predictors. a survey, ACM Computing Surveys 58 (2025). URL: https://doi.org/10.1145/3749097. doi:10.1145/3749097.

[11] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, Symbolic knowledge extraction from opaque ML predictors in PSyKE: Platform design & experiments, Intelligenza Artificiale 16 (2022) 27–48. doi:10.3233/IA-210120.

[12] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, On the design of PSyKE: A platform for symbolic knowledge extraction, in: R. Calegari, G. Ciatto, E. Denti, A. Omicini, G. Sartor (Eds.), WOA 2021 – 22nd Workshop "From Objects to Agents", volume 2963 of *CEUR Workshop Proceedings*, Sun SITE Central Europe, RWTH Aachen University, 2021, pp. 29–48. URL: https://ceur-ws.org/Vol-2963/paper14.pdf, 22nd Workshop "From Objects to Agents" (WOA 2021), Bologna, Italy, September 1–3, 2021. Proceedings.

[13] R. Calegari, F. Sabbatini, The PSyKE technology for trustworthy artificial intelligence, in: A. Dovier, A. Montanari, A. Orlandini (Eds.), AIxIA 2022, volume 13796 of *Lecture Notes in Computer Science*, Springer, Cham, Switzerland, 2023, pp. 3–16. doi:10.1007/978-3-031-27181-6_1, XXI International Conference of the Italian Association for Artificial Intelligence, AIxIA 2022, Udine, Italy, November 28 – December 2, 2022, Proceedings.

[14] R. Calegari, G. G. Castañé, M. Milano, B. O'Sullivan, et al., Assessing and enforcing fairness in the ai lifecycle, in: IJCAI, volume 2023, International Joint Conferences on Artificial Intelligence, 2023, pp. 6554–6562.

[15] J. Chai, T. Jang, X. Wang, Fairness without demographics through knowledge distillation, Advances in Neural Information Processing Systems 35 (2022) 19152–19164.

[16] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215. doi:10.1038/s42256-019-0048-x.

[17] Z. Zhu, S. Si, J. Wang, Y. Yang, J. Xiao, Debias the black-box: A fair ranking framework via knowledge distillation, in: International Conference on Web Information Systems Engineering, Springer, 2022, pp. 395–405.

[18] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, Classification and Regression Trees, CRC Press, 1984.

[19] J. Huysmans, B. Baesens, J. Vanthienen, ITER: An algorithm for predictive regression rule extraction, in: Data Warehousing and Knowledge Discovery (DaWaK 2006), Springer, 2006, pp. 270–279. doi:10.1007/11823728_26.

[20] F. Sabbatini, G. Ciatto, A. Omicini, GridEx: An algorithm for knowledge extraction from black-box regressors, in: Explainable and Transparent AI and Multi-Agent Systems. Third International Workshop, EXTRAAMAS 2021, Virtual Event, May 3–7, 2021, Revised Selected Papers, volume 12688 of *LNCS*, Springer Nature, Basel, Switzerland, 2021, pp. 18–38. doi:10.1007/978-3-030-82017-6_2.

[21] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, Towards a unified model for symbolic knowledge extraction with hypercube-Based methods, Intelligenza Artificiale 17 (2023) 63–75. doi:10.3233/IA-230001.

[22] F. Sabbatini, R. Calegari, Unveiling opaque predictors via explainable clustering: The CReEPy algorithm, in: G. Boella, F. A. D'Asaro, A. Dyoub, L. Gorrieri, F. A. Lisi, C. Manganini, G. Primiero (Eds.), Proceedings of the 2nd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023), Rome, Italy, November 6, 2023, volume 3615 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 1–14. URL: https://ceur-ws.org/Vol-3615/paper1.pdf.

[23] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, Hypercube-Based methods for symbolic knowledge extraction: Towards a unified model, in: A. Ferrando, V. Mascardi (Eds.), WOA 2022 – 23rd Workshop "From Objects to Agents", volume 3261 of *CEUR Workshop Proceedings*, Sun SITE Central Europe, RWTH Aachen University, 2022, pp. 48–60. URL: http://ceur-ws.org/Vol-3261/paper4.pdf.

[24] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10–13, 2015, ACM, 2015, pp. 259–268. URL: https://doi.org/10.1145/2783258.2783311. doi:10.1145/2783258.2783311.

[25] M. J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual fairness, in: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.