

# Unmasking the Shadows: Leveraging Symbolic Knowledge Extraction to Discover Biases and Unfairness in Opaque Predictive Models

Federico Sabbatini<sup>1,\*</sup>, Roberta Calegari<sup>2</sup>

<sup>1</sup>University of Urbino Carlo Bo

<sup>2</sup>Alma Mater Studiorum—University of Bologna

## Abstract

This work explores the efficacy of symbolic knowledge-extraction (SKE) techniques in identifying biases and unfairness within opaque predictive models. Logic rules extracted from black-box predictors make it possible to verify if decisions are influenced by protected or sensitive features. In particular, the identification of biased or unfair decisions can be achieved through the evaluation of *if-then* rules, detecting the inclusion of protected and/or sensitive information in the rules' precondition. The effectiveness of SKE in this regard is demonstrated here by conducting various simulations on a well-known data set for loan grant prediction. Our findings highlight the potential of SKE as a valuable tool to reveal biases and discrimination in opaque predictions, ultimately contributing to the pursuit of fair and transparent decision-making systems.

## Keywords

Fairness in AI, Bias in AI, Explainable artificial intelligence, XAI, Symbolic knowledge extraction, PSyKE

## 1. Introduction

As predictive models become increasingly integrated into various domains, ensuring their fairness and transparency is of paramount importance [1]. Opaque predictive models in machine learning (ML), often referred to as black-box models, pose challenges in understanding the underlying mechanisms by which they make predictions. Consequently, biases and discrimination can inadvertently permeate these models, leading to unfair or prejudiced outcomes [2]. To address this critical issue, the present paper investigates the application of symbolic knowledge-extraction (SKE) techniques in uncovering biases and discrimination within opaque predictive models.

SKE offers a promising avenue to extract interpretable logic rules from black-box models, enabling a deeper understanding of decision-making [3, 4]. By distilling complex model behaviours into human-readable rules, SKE facilitates the identification of specific conditions under which biases may arise. This approach proves particularly valuable when examining whether protected features play a role in decision-making since the presence of protected information in the

---

AEQUITAS 2024: Workshop on Fairness and Bias in AI | co-located with ECAI 2024, Santiago de Compostela, Spain


\*Corresponding author.

✉ f.sabbatini1@campus.uniurb.it (F. Sabbatini); roberta.calegari@unibo.it (R. Calegari)

🆔 0000-0002-0532-6777 (F. Sabbatini); 0000-0003-3794-2942 (R. Calegari)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

preconditions of extracted rules can provide direct evidence of bias. The same considerations may also extend to sensitive features, e.g., those that are not protected themselves but are related to other features identified as protected (e.g., name or height, which allow ML models to infer race and/or gender of individuals; [5, 6, 7]). We point out that identifying correlations between protected/sensitive features and other input variables is not within the scope of SKE techniques, nor is the recognition of protected/sensitive attributes in the rule preconditions<sup>1</sup>. The classification of input features into “unfairness-enablers” and “potentially-fairness-neutral” should be performed by human users as an independent task.

The main objective of this paper is to demonstrate the effectiveness of SKE in identifying unfairness and discrimination within opaque predictions. To achieve this, we employ a well-known classification data set aimed at predicting loan grants. We conduct various simulations to illustrate how SKE can be exploited to extract logic rules and evaluate their fairness implications. Through these examples, we aim to shed light on the potential of SKE as a practical tool for highlighting biases and promoting fairness in predictive modelling.

By revealing biases and discrimination present in opaque predictive models, this research contributes to the broader discourse on fairness, accountability, and transparency in algorithmic decision-making. Understanding and rectifying biases in these models are crucial steps towards building equitable systems that mitigate the perpetuation of societal inequalities. The insights gained from this study serve as a foundation for developing strategies to enhance fairness in predictive models and promote the responsible deployment of artificial intelligence (AI) in critical domains.

In the following sections, we will discuss the methodology employed for SKE, present the results of our experiments, and discuss the implications and potential future directions of this research. By critically examining the power of SKE in identifying biases, we hope to provide practical insights and actionable recommendations for researchers, practitioners, and policymakers working towards fair and transparent predictive models.

## 2. Related Works

Several studies have explored different approaches and methodologies to address bias in AI.

One line of research focuses on rule-based techniques for bias detection and explanation [8, 9]. These studies aim to extract interpretable rules from black-box models and analyse them for potential biases. For instance, in [8] the authors have proposed algorithms mining association rules or decision trees to identify discriminatory patterns in the rule sets generated by predictive models. These approaches often leverage fairness criteria or sensitive attribute definitions to guide the rule extraction process.

Another area of related work involves the use of fairness-aware machine learning techniques [10, 11]. These approaches aim to incorporate fairness considerations during the model training phase, ensuring that the resulting predictions are less likely to be biased. Fairness-aware algorithms often employ mathematical optimisation techniques to balance predictive accuracy

---

<sup>1</sup>In the following, we adopt the terms “protected” and “sensitive” as synonyms, since the considerations discussed in this work apply to both categories

and fairness objectives, taking into account various fairness definitions such as demographic parity [12], equalised odds [13], or individual fairness [14].

Furthermore, researchers have explored post-hoc methods to detect and mitigate biases in predictive models [15, 16]. These methods involve analysing the outcomes of model predictions on different subgroups defined by sensitive attributes, such as race, gender, or age. By quantifying and comparing the disparities in prediction outcomes across subgroups, these techniques can help identify and address discriminatory behaviour in models.

SKE techniques, including rule extraction and logic rule analysis, have been used in various domains to interpret and understand black-box models [17, 18, 19, 20]. However, thus far, their specific application for bias and discrimination identification in opaque predictions has not gained much attention. The proposed research aims to contribute to this body of work by demonstrating the effectiveness of SKE in uncovering biases and discrimination and providing insights into its practical application for fairness assessment in predictive models.

Through a comprehensive review of existing related work, this paper will situate SKE methods within the broader context of bias detection and fairness assessment in predictive modelling. It will build upon and extend the current knowledge by showcasing the unique capabilities of SKE techniques in addressing biases and discrimination in opaque predictive models, thus contributing to the growing literature on fair and transparent algorithmic decision-making.

### 3. Symbolic Knowledge Extraction: Methods and Methodology

SKE is a methodology aiming to extract interpretable and logic rules from complex black-box models, enabling a deeper understanding of their decision-making processes. There are two main approaches within SKE: pedagogical and decompositional [21].

In the pedagogical approach, the focus is on extracting human-readable rules providing meaningful explanations of the model’s behaviour. These rules are often represented in *if-then* format, making them easily understandable by both humans and machines. The pedagogical approach prioritises generality, allowing stakeholders to gain insights into the decision criteria employed by any predictive model, even though the explanations may lose some of the underlying model’s complexity and performance.

On the other hand, the decompositional approach aims to decompose the black-box model into simpler, more interpretable sub-models or components that are typically easier to understand and analyse individually. The inner black-box structure is carefully analysed and the resulting explanations may be more adherent to the underlying model behaviour. However, these techniques are strictly tailored to narrow categories of predictors, thus lacking flexibility and generality.

Since both approaches generate intuitive explanations that can be easily communicated and understood by a broader audience, this work prioritises bias evaluations independent of the underlying predictive model. Therefore, we exploit pedagogical approaches as the main tools for our experiments.

In the following, we provide a summary of some state-of-the-art pedagogical SKE techniques – namely, GRIDEx, CART and CREPY – offering insights into the specific techniques employed in the experimentation section.

We leverage the implementations available within the PSyKE Python package<sup>2</sup> [22, 23]. This library encompasses all the aforementioned SKE implementations, allowing for their seamless comparison and evaluation [24]. The PSyKE platform offers a unified interface, enabling the application, assessment, and comparison of various SKE techniques. Moreover, it is fully compatible with other widely-used Python packages [25], such as Scikit-Learn [26], and provides additional extensions for SKE [27] and functionalities for feature engineering, data manipulation and visualisation, Semantic Web compatibility [28], and assessment of knowledge quality [29, 30].

### 3.1. GRIDEx

GRIDEx [31] is a pedagogical SKE algorithm originally designed for regression tasks and based on hypercubic partitioning of the input feature space. The partitioning is recursive, symmetric and performed top-down to obtain human-interpretable rules describing as many disjoint, hypercubic input space subregions. Thanks to the generalisation presented in [32, 33], it is possible to apply GRIDEx to both classification and regression tasks if they are encoded via data sets having only continuous input features.

GRIDEx requires the following set of hyper-parameters to be defined by users:

**recursion depth** defining the maximum number of recursions to perform during the knowledge extraction;

**splitting strategy** to partition the input space. It may be fixed, if each input dimension is split into a fixed number of partitions, or adaptive if the number of splits depends on the relevance of the features;

**number of splits** defining how many slices have to be performed along each input dimension;

**error threshold** used to decide on which regions the recursive step of the algorithm has to be performed. In particular, only regions with a predictive error greater than the user-defined threshold are recursively split.

This set of parameters may be automatically tuned with the PEDRO procedure [34].

### 3.2. CART

The CART algorithm [35] is based on the induction of a classification or regression binary decision tree. It may be directly applied to a data set to build a human-interpretable predictor (if the induced tree is not deep) or it may be adopted as an SKE technique to produce human-interpretable rules mimicking the behaviour of an opaque ML model. Human-interpretable rules are obtained by reading the complete paths from the tree root to each distinct leaf, given that internal nodes represent constraints on input variables and leaves contain output predictions.

The most important parameters to consider for CART are:

**maximum depth** defining the maximum allowed depth for the decision tree;

---

<sup>2</sup><https://github.com/psykei/psyke-python>

**maximum number of leaves** defining the maximum allowed number of tree leaves.

These two parameters are intertwined and both the predictive accuracy and the human-readability extent of the tree critically depend on them. In particular, deep trees usually exhibit higher predictive performance but smaller human-readability extent than shallow ones. The same holds for trees with a large number of leaves compared to trees with fewer leaves.

### 3.3. CREEPY

The CREEPY algorithm [36, 37] is a pedagogical SKE technique applicable to opaque classifiers and regressors. It relies on underlying explainable clustering procedures aimed at identifying hypercubic human-interpretable regions within the input feature space [38, 39]. At the end of the knowledge extraction, each hypercubic region is translated into a Prolog rule describing the boundaries of the region and the corresponding output prediction.

Suitable explainable clusterings adopted by CREEPY are CREAM [40] and ExACT [41]. They both perform hierarchical clustering according to different recursive strategies and require the following parameters, possibly tuned with the ORCHiD automated procedure [40]:

**recursion depth** defining the maximum number of performed recursions;

**maximum number of Gaussian components** defining the maximum number of components to use in the Gaussian mixture model clustering performed within ExACT and CREAM;

**error threshold** used to pre-emptively stop the recursive clustering when clusters exhibit a predictive error smaller than the threshold.

To execute CREEPY users have to provide the parameters required by the underlying instance of ExACT or CREAM as well as an optional feature relevance threshold used to drop from the output Prolog rules all the antecedents involving input features with relevance below the threshold.

## 4. Experiments

### 4.1. Running Example: the Loan Data Set Case Study

We selected the Loan data set<sup>3</sup> as a case study to carry out experiments and verify if SKE techniques are effective tools to identify discriminative predictions provided by opaque models. The data set is composed of 11 input features representing relevant variables to decide if a loan should be granted or not. The final decision is the binary output feature. The data set is completed by an additional feature representing a unique identification code for each loan. The data set counts 614 instances. Only 480 have no missing values. The names, types, and values of the features are reported in Table 1.

---

<sup>3</sup><https://www.kaggle.com/datasets/burak3ergun/loan-data-set>

**Table 1**  
Loan data set features.

Feature name	Type	Values
Gender	Binary, nominal	Female, Male
Married	Binary, nominal	No, Yes
Dependents	Discrete, nominal	0, 1, 2, 3+
Education	Binary, nominal	Graduate, Not graduate
Self employed	Binary, nominal	No, Yes
Applicant income	Numeric	from 150 to 81000
Coapplicant income	Numeric	from 0 to 33837
Loan amount	Numeric	from 9 to 600
Loan amount term	Discrete, numeric	9 distinct values between 36 and 480
Credit history	Binary, numeric	0, 1
Property area	Discrete, nominal	Rural, Semiurban, Urban
Loan status	Binary, nominal	No, Yes

In conducting the experiments presented in this study, instances in the data set that contained missing values were excluded, and nominal attributes were converted into discrete numeric features.

To evaluate the fairness of the data sets and opaque predictors, we employed the disparate impact index [42]. This metric measures the extent of differential treatment between two distinct groups, specifically by quantifying the proportion of individuals from each group who receive positive outcomes. The disparate impact index serves as a quantitative measure of the disparate treatment experienced by individuals from different classes.

The calculation of the disparate impact index involves grouping the instances in a data set  $\mathcal{S}$  into two subgroups: a privileged (or base) group  $\mathcal{S}^P$  and an unprivileged (or protected) group  $\mathcal{S}^U$ , typically affected by fairness concerns. Formally,

$$\mathcal{S} = \{ \mathbf{x}_i : \mathbf{x}_i = (x_i^1, x_i^2, \dots, x_i^d) \},$$

where  $x_i^1, x_i^2, \dots, x_i^d$  are the  $d$  features of instance  $\mathbf{x}_i$  and

$$\mathcal{S}^P = \{ \mathbf{x}_i : \mathbf{x}_i \in \mathcal{S} \wedge x_i^\pi = \oplus \},$$

$$\mathcal{S}^U = \{ \mathbf{x}_i : \mathbf{x}_i \in \mathcal{S} \wedge x_i^\pi = \ominus \},$$

by assuming that the sensitive feature  $\pi$  have values in  $\{\oplus, \ominus\}$ , with  $x^\pi = \oplus$  representing the membership to the privileged group.

For each group, the ratio of positive outcomes to the total number of individuals is computed. Subsequently, the disparate impact index, denoted as  $DI$ , is defined as follows:

$$DI = \frac{\frac{|\{ \mathbf{x}_i : \mathbf{x}_i \in \mathcal{S}^U \wedge \gamma(\mathbf{x}_i) = \odot \}|}{|\mathcal{S}^U|}}{\frac{|\{ \mathbf{x}_i : \mathbf{x}_i \in \mathcal{S}^P \wedge \gamma(\mathbf{x}_i) = \odot \}|}{|\mathcal{S}^P|}}, \quad (1)$$

**Table 2**

DI scores calculated for the Loan data set and the corresponding predictions generated by RF classifiers. \* denotes “any possible value”.

Loan outcome	Male			Female			DI index
	*	Yes	No	*	Yes	No	
Data set (original)	394	278	116	86	54	32	0.890
Data set (28% perturbed)	394	278	116	86	39	47	0.643
Data set (56% perturbed)	394	278	116	86	24	62	0.396
Data set (83% perturbed)	394	278	116	86	9	77	0.148
RF (original) (accuracy = 0.79)	394	333	61	86	72	14	0.991
RF (28% perturbed) (accuracy = 0.75)	394	336	58	86	58	28	0.791
RF (56% perturbed) (accuracy = 0.79)	394	336	58	86	5	81	0.068
RF (83% perturbed) (accuracy = 0.79)	394	335	59	86	1	85	0.014

where  $\gamma(\mathbf{x}_i)$  represents the output of instance  $\mathbf{x}_i$  and  $\odot$  is the positive output.

In our experimental setup, we specifically focus on the scenario of gender discrimination ( $\pi = \text{gender}$ ). Consequently, we designate female individuals as the unprivileged group ( $\ominus = \text{female}$ ) and male individuals as the privileged group ( $\oplus = \text{male}$ ). This choice allows us to investigate and analyse potential biases and disparities that may affect females within the context of the studied predictive models. In our case study  $\gamma$  is a function denoting the approval or denial of a loan (therefore,  $\text{loan}(\mathbf{x}) = \text{yes}$  corresponds to a positive outcome). The *DI* is accordingly defined as follows:

$$DI = \frac{\left| \left\{ \mathbf{x}_i : \mathbf{x}_i \in \mathcal{S} \wedge x_i^{\text{gender}} = \text{female} \wedge \text{loan}(\mathbf{x}_i) = \text{yes} \right\} \right|}{\frac{\left| \left\{ \mathbf{x}_i : \mathbf{x}_i \in \mathcal{S} \wedge x_i^{\text{gender}} = \text{female} \right\} \right| \cdot \left| \left\{ \mathbf{x}_i : \mathbf{x}_i \in \mathcal{S} \wedge x_i^{\text{gender}} = \text{male} \wedge \text{loan}(\mathbf{x}_i) = \text{yes} \right\} \right|}{\left| \left\{ \mathbf{x}_i : \mathbf{x}_i \in \mathcal{S} \wedge x_i^{\text{gender}} = \text{male} \right\} \right|}}. \quad (2)$$

As reported in the first row of Table 2, 394 out of 480 instances describe loans demanded by male applicants. The remaining 86 instances correspond to female applicants. Even if the gender attribute is not balanced, it is possible to observe that loans are fairly granted to female and male applicants. Indeed, 278 out of 394 male applicants receive the loan, as well as 54 out of 86 female applicants. This corresponds to the 71% and 63% of male and female applicants, respectively. By applying Equation (2) it is possible to find a disparate impact score of 0.89, corresponding to a quite fair situation. We recall here that  $DI = 1$  denotes a perfectly fair situation. Lower score values are associated with unfair conditions. A score of 0.8 is usually considered the threshold to divide fairness ( $DI > 0.8$ ) from unfairness ( $DI < 0.8$ ). As a consequence of all these observations, we consider the Loan data set fair from the gender standpoint.

The distribution of the output feature of the data set with respect to the gender attribute is visually presented in Figure 1a. The x-axis represents the credit history input feature, which is considered the most significant for classification purposes. Gender is reported in the y-axis. The



**Table 3**

Parameters adopted to perform knowledge extraction with CART, GRIDEx and CREEPY from the RF classifiers trained on the Loan data set.

Extractor	Parameters
CART	Maximum depth = 2 Maximum leaf amount = unbounded
GRIDEx	Maximum recursion depth = 1 Splitting strategy = adaptive Spits = 3 along the most relevant input feature 2 along the second most relevant input feature 1 along the other input features Error threshold = 0.1
CREEPY	Underlying clustering = CREAM Maximum recursion depth = 3 Maximum Gaussian components = 2 Error threshold = 0.01

size of the circles corresponds to the number of instances in each subregion of the input feature space. Orange circles indicate granted loans, whereas green circles indicate denied loans.

#### 4.2. SKE on the Loan Data Set: Uncovering Insights and Patterns

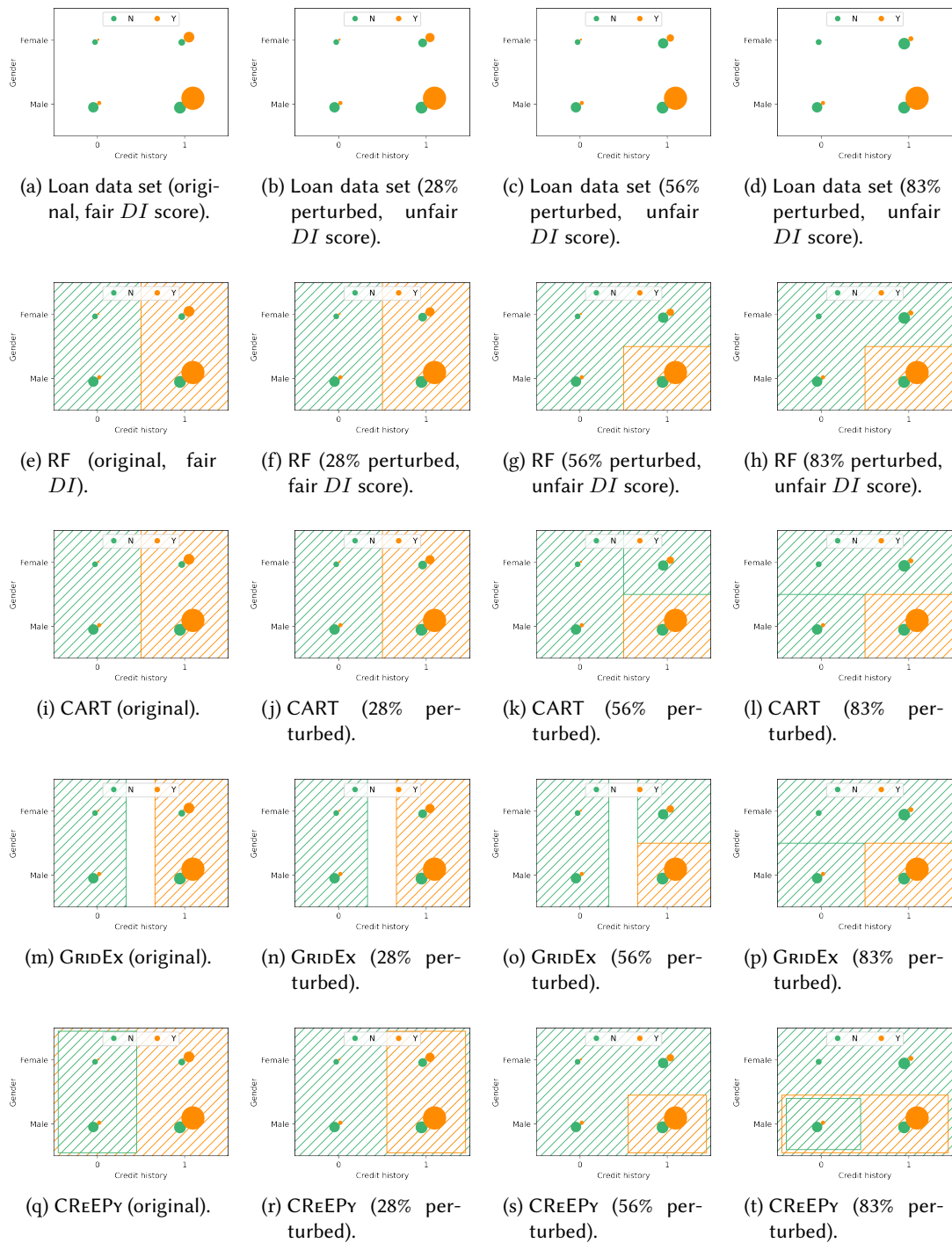
A random forest (RF) classifier has been trained upon the Loan data set. The data set has been randomly split into training (85%) and test (15%) sets. The RF predictor was composed of 50 base decision trees having a maximum depth of 5 and achieved a classification accuracy equal to 0.79. The decision boundaries of the RF classifier are reported in Figure 1e as a bidimensional projection on the credit history and gender input features.

The RF can be considered a fair predictor since its disparate impact score is equal to 0.99 (cf. first row of the bottom part of Table 2). It is worth mentioning that fairness is not directly associated with classification accuracy. In this particular case, despite the RF classifier's predictive performance not being excellent, it is noteworthy that it demonstrates a high level of fairness from a gender perspective. Fairness, in this context, refers to the absence of bias or discrimination based on gender, regardless of the classifier's overall accuracy in making predictions.

The goal of our experiments is to demonstrate if SKE techniques can be used to detect unfair opaque predictors. To this purpose, we use the CART, GRIDEx and CREEPY algorithms to perform knowledge extraction on the trained RF classifier. Extractors have been parametrised as summarised in Table 3. The number of extracted rules, as a proxy of the human-readability extent of the models, and the fidelity measured for each extractor with respect to the RF predictions, expressed as classification accuracy, have been reported in Table 4. All extractors can achieve a fidelity of 0.99 with 2 rules.

The decision boundaries obtained via CART, GRIDEx and CREEPY are reported in Figures 1i, 1m and 1q, respectively. The corresponding Prolog rules are shown in Listings 1 to 3, respectively.





**Figure 1:** Visualisation of loan data set output distribution with respect to the most relevant input feature (i.e., credit history) and the gender feature. The circle sizes represent the number of instances for each input coordinate pair. Decision boundaries are illustrated for an RF opaque predictor and various SKE techniques. Columns progressively demonstrate increasing bias and discrimination, indicated by a greater number of loans denied to female applicants.

**Table 4**

Predictive performance and human-readability extent of all SKE techniques applied to the described RF classifiers.

Opaque predictor	Extractor	Fidelity	Extracted rules
RF (original)	CART	0.99	2
	GRIDEx	0.99	2
	CRÉEPY	0.99	2
RF (28% perturbed)	CART	0.99	2
	GRIDEx	0.99	2
	CRÉEPY	0.99	2
RF (56% perturbed)	CART	0.97	3
	GRIDEx	0.97	3
	CRÉEPY	0.97	2
RF (83% perturbed)	CART	1.00	3
	GRIDEx	1.00	3
	CRÉEPY	1.00	3

---

Listing 1: Rules extracted with CART for the Loan data set (original and 28% perturbed data set).

```
loan(Gender, Married, ..., CreditHistory, PropertyArea, NO) :-
    CreditHistory < 0.5.
loan(Gender, Married, ..., CreditHistory, PropertyArea, YES).
```

---



---

Listing 2: Rules extracted with GRIDEx for the Loan data set (original and 28% perturbed data set).

```
loan(Gender, Married, ..., CreditHistory, PropertyArea, NO) :-
    CreditHistory in [0.00, 0.33].
loan(Gender, Married, ..., CreditHistory, PropertyArea, YES) :-
    CreditHistory in [0.67, 1.00].
```

---



---

Listing 3: Rules extracted with CRÉEPY for the Loan data set (original).

```
loan(Gender, Married, ..., CreditHistory, PropertyArea, NO) :-
    CreditHistory in [0.00, 0.00].
loan(Gender, Married, ..., CreditHistory, PropertyArea, YES).
```

---

The three SKE algorithms reveal that the predictions made by the RF model are solely influenced by the credit history input feature. Irrespective of the applicants' gender, loans are granted to individuals with a positive credit history (credit history = 1), while they are denied to those with a negative credit history (credit history = 0). The SKE techniques confirm the RF's fair behaviour concerning the applicants' gender, as the predictions are solely driven by the credit history attribute and are independent on gender.

---

Listing 4: Rules extracted with CREEPY for the Loan data set (28% perturbed).

---

```
loan(Gender, Married, ..., CreditHistory, PropertyArea, YES) :-  
    CreditHistory in [1.00, 1.00].  
loan(Gender, Married, ..., CreditHistory, PropertyArea, NO).
```

---

### 4.3. Injecting Bias in the Loan Data Set

To inject bias, we perturbed the output feature of the Loan data set, which was originally fair with respect to gender. The perturbation involved changing the loan status from ‘Yes’ to ‘No’ for a variable number of female applicants. Specifically, we conducted three different perturbations, modifying the positive loan outcome for 15, 30, and 45 female applicants. These numbers correspond to 28%, 56%, and 83% respectively, of the total female applicants who originally had a positive loan outcome in the unaltered data set.

The output feature distribution for the biased data sets is reported in Figures 1b to 1d and in the top part of Table 2. The corresponding disparate impact measurements are reported in the rightmost column of the same table. As expected, the score values decrease by increasing the introduced bias, down to 0.15 for the most perturbed data set. Each data set has been used to train an RF classifier with 50 base predictors having maximum depth equal to 5 and a measured predictive accuracy on the test set varying between 0.75 and 0.79 (cf. bottom part of Table 2).

#### 4.3.1. 28% Perturbed Data Set

The RF classifier trained upon the Loan data set with a perturbation involving 28% of the positive female applicants has  $DI = 0.79$ , even though the biased data set has a lower score ( $DI = 0.64$ ). This difference is due to the predictive error of the RF. There are no noticeable differences in the decision boundaries of this RF compared to the one trained on the original Loan data set (cf. Figures 1e and 1f). Also CART, GRIDEX and CREEPY applied to the RF provide outputs similar to those obtained for the unbiased case study (see Figures 1j, 1n and 1r). The only difference is the Prolog theory obtained via CREEPY, having the same semantics as the unbiased counterpart, but different clauses. The theory is listed in Listing 4.

Also in this case the human-interpretable rules extracted via SKE techniques do not identify discriminative predictions based on gender for the RF classifier and this is in agreement with the corresponding disparate impact scores.

#### 4.3.2. 56% Perturbed Data Set

A different situation is evident if we modify the data set in order to refuse the loan to the 56% of female applicants that conversely should have received it. In this case, the disparate impact score drops to 0.40 for the data set and to 0.07 for the corresponding trained RF. These values highlight strong unfairness, especially for the RF predictions. The corresponding decision boundaries are reported in Figure 1g. It is clearly visible that the loan is granted only to male applicants having a positive credit history.

Decision boundaries obtained via CART, GRIDEX and CREEPY and the corresponding extracted rules expressed as Prolog theories are reported in Figures 1k, 1o and 1s and Listings 5 to 7,

---

Listing 5: Rules extracted with CART for the Loan data set (56% perturbed).

---

```
loan(Gender, Married, ..., CreditHistory, PropertyArea, NO) :-  
    CreditHistory < 0.5.  
loan(Gender, Married, ..., CreditHistory, PropertyArea, NO) :-  
    Gender = 'Female'.  
loan(Gender, Married, ..., CreditHistory, PropertyArea, YES).
```

---

---

Listing 6: Rules extracted with GRiDEX for the Loan data set (56% perturbed).

---

```
loan(Gender, Married, ..., CreditHistory, PropertyArea, NO) :-  
    CreditHistory in [0.00, 0.33].  
loan(Gender, Married, ..., CreditHistory, PropertyArea, YES) :-  
    CreditHistory in [0.67, 1.00], Gender in ['Male'].  
loan(Gender, Married, ..., CreditHistory, PropertyArea, NO) :-  
    CreditHistory in [0.67, 1.00], Gender in ['Female'].
```

---

---

Listing 7: Rules extracted with CRiEPY for the Loan data set (56% perturbed).

---

```
loan(Gender, Married, ..., CreditHistory, PropertyArea, YES) :-  
    CreditHistory in [0.00, 0.00], Gender in ['Male'].  
loan(Gender, Married, ..., CreditHistory, PropertyArea, NO).
```

---

---

Listing 8: Rules extracted with CART for the Loan data set (83% perturbed).

---

```
loan(Gender, Married, ..., CreditHistory, PropertyArea, NO) :-  
    Gender = 'Female'.  
loan(Gender, Married, ..., CreditHistory, PropertyArea, NO) :-  
    CreditHistory < 0.5.  
loan(Gender, Married, ..., CreditHistory, PropertyArea, YES).
```

---

respectively.

In this scenario, the credit history of the applicant remains the primary feature considered during the prediction phase of the RF model. For instance, the initial split in the input feature space performed by CART focuses on this attribute. However, the gender feature also plays a role in predicting the outcomes for a subset of instances, specifically those with a good credit history. As a result, the SKE techniques demonstrate their effectiveness in identifying unfair predictors by revealing the influence of the gender attribute on outcomes within specific credit history subgroups.

#### 4.3.3. 83% Perturbed Data Set

Finally, we report here the results obtained for the Loan data set with a perturbation involving 83% of the female applicants receiving positive outcomes. The disparate impact scores for this data set and the corresponding trained RF are equal to 0.15 and 0.01, respectively. The scores highlight severe unfairness. Decision boundaries identified by the RF, CART, GRiDEX

---

Listing 9: Rules extracted with GRIDEx for the Loan data set (83% perturbed).

---

```
loan(Gender, Married, ..., CreditHistory, PropertyArea, NO) :-  
    CreditHistory in [0.00, 1.00], Gender in ['Female'].  
loan(Gender, Married, ..., CreditHistory, PropertyArea, NO) :-  
    CreditHistory in [0.00, 0.50], Gender in ['Male'].  
loan(Gender, Married, ..., CreditHistory, PropertyArea, YES) :-  
    CreditHistory in [0.50, 1.00], Gender in ['Male'].
```

---

Listing 10: Rules extracted with CRÉPY for the Loan data set (83% perturbed).

---

```
loan(Gender, Married, ..., CreditHistory, PropertyArea, NO) :-  
    CreditHistory in [0.00, 0.50], Gender in ['Male'].  
loan(Gender, Married, ..., CreditHistory, PropertyArea, YES) :-  
    CreditHistory in [0.50, 1.00], Gender in ['Male'].  
loan(Gender, Married, ..., CreditHistory, PropertyArea, NO).
```

---

and CRÉPY are reported in Figures 1h, 1l, 1p and 1t, respectively. Prolog rules provided by the SKE techniques are reported in Listings 8 to 10.

The extracted rules clearly emphasise the significant reliance of the RF predictions on the gender feature. Despite the decision boundaries being the same as in the previous case study with a 56% perturbation, in this instance gender is employed as the primary feature for decision-making, followed by credit history as the secondary feature. Essentially, loans are primarily granted or denied based on gender, with credit history playing a secondary role. The SKE techniques effectively identify and reveal this unfair behaviour, presenting it to human users in the form of interpretable logic rules.

## 5. Conclusion

This paper provides preliminary insights into the value of leveraging SKE techniques for studying biases in AI predictors. The findings demonstrate the potential of SKE techniques, particularly in analysing the relationships between decision outcomes and sensitive input attributes. This work highlights the importance of considering the correlation between decisions and sensitive attributes, such as gender, and how SKE can effectively identify and highlight these dependencies.

Looking ahead, future research will focus on further testing and refining the proposed approach. This will involve exploring the application of SKE techniques with proxy variables, investigating intersectional discrimination, and employing counterfactual techniques. Additionally, the study will delve into the evaluation of different fairness metrics to gain a more comprehensive understanding of bias and discrimination within predictive models.

Merging the field of AI fairness with explainable AI seems to be a promising approach. By doing so, we can develop robust methodologies to mitigate biases and promote fairness in AI systems. The ongoing exploration of SKE techniques holds great promise in fostering a more equitable and unbiased landscape for AI decision-making.

## Acknowledgments

This work has been supported by the EU ICT-48 2020 project TAILOR (No. 952215) and the European Union's Horizon Europe AEQUITAS research and innovation programme under grant number 101070363.

## References

- [1] S. Vollmer, B. A. Mateen, G. Bohner, F. J. Király, R. Ghani, P. Jonsson, S. Cumbers, A. Jonas, K. S. McAllister, P. Myles, et al., Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness, *bmj* 368 (2020).
- [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] R. Calegari, G. Ciatto, A. Omicini, On the integration of symbolic and sub-symbolic techniques for XAI: A survey, *Intelligenza Artificiale* 14 (2020) 7–32. doi:10.3233/IA-190036.
- [4] 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. URL: <https://doi.org/10.1145/3645103>. doi:10.1145/3645103.
- [5] T. L. Quy, A. Roy, V. Iosifidis, W. Zhang, E. Ntoutsis, A survey on datasets for fairness-aware machine learning, *WIREs Data Mining and Knowledge Discovery* 12 (2022). URL: <https://doi.org/10.1002/widm.1452>. doi:10.1002/WIDM.1452.
- [6] T. van Nuenen, X. Ferrer, J. M. Such, M. Coté, Transparency for whom? assessing discriminatory artificial intelligence, *Computer* 53 (2020) 36–44. URL: <https://doi.org/10.1109/MC.2020.3002181>. doi:10.1109/MC.2020.3002181.
- [7] B. Wiggins, *Calculating Race: Racial Discrimination in Risk Assessment*, Oxford University Press, 2020. URL: <https://doi.org/10.1093/oso/9780197504000.001.0001>. doi:10.1093/oso/9780197504000.001.0001.
- [8] T. Calders, S. Verwer, Three naive bayes approaches for discrimination-free classification, *Data mining and knowledge discovery* 21 (2010) 277–292.
- [9] D. Pedreshi, S. Ruggieri, F. Turini, Discrimination-aware data mining, in: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08*, Association for Computing Machinery, New York, NY, USA, 2008, p. 560–568. URL: <https://doi.org/10.1145/1401890.1401959>. doi:10.1145/1401890.1401959.
- [10] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning fair representations, in: *International conference on machine learning*, PMLR, 2013, pp. 325–333.
- [11] T. Kamishima, S. Akaho, H. Asoh, J. Sakuma, Fairness-aware classifier with prejudice remover regularizer, in: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012*, Bristol, UK, September 24–28, 2012. *Proceedings, Part II 23*, Springer, 2012, pp. 35–50.
- [12] M. Hort, Z. Chen, J. M. Zhang, F. Sarro, M. Harman, Bias mitigation for machine learning

- classifiers: A comprehensive survey, CoRR abs/2207.07068 (2022). URL: <https://doi.org/10.48550/arXiv.2207.07068>. doi:10.48550/ARXIV.2207.07068. arXiv:2207.07068.
- [13] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, December 5–10, 2016, Barcelona, Spain, 2016, pp. 3315–3323. URL: <https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html>.
- [14] J. Chakraborty, K. Peng, T. Menzies, Making fair ML software using trustworthy explanation, in: *35th IEEE/ACM International Conference on Automated Software Engineering, ASE 2020*, Melbourne, Australia, September 21–25, 2020, IEEE, 2020, pp. 1229–1233. URL: <https://doi.org/10.1145/3324884.3418932>. doi:10.1145/3324884.3418932.
- [15] D. Madras, T. Pitassi, R. Zemel, Predict responsibly: improving fairness and accuracy by learning to defer, *Advances in Neural Information Processing Systems 31* (2018).
- [16] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.
- [17] G. Bologna, C. Pellegrini, Three medical examples in neural network rule extraction, *Physica Medica* 13 (1997) 183–187. URL: <https://archive-ouverte.unige.ch/unige:121360>.
- [18] L. Franco, J. L. Subirats, I. Molina, E. Alba, J. M. Jerez, Early breast cancer prognosis prediction and rule extraction using a new constructive neural network algorithm, in: *Computational and Ambient Intelligence (IWANN 2007)*, volume 4507 of *LNCS*, Springer, 2007, pp. 1004–1011. doi:0.1007/978-3-540-73007-1\_121.
- [19] R. Setiono, B. Baesens, C. Mues, Rule extraction from minimal neural networks for credit card screening, *International Journal of Neural Systems* 21 (2011) 265–276. doi:10.1142/S0129065711002821.
- [20] F. Sabbatini, C. Grimani, R. Calegari, Bridging machine learning and diagnostics of the esa lisa space mission with equation discovery via explainable artificial intelligence, *Advances in Space Research* 74 (2024) 505–517. URL: <https://www.sciencedirect.com/science/article/pii/S0273117724003880>. doi:<https://doi.org/10.1016/j.asr.2024.04.041>.
- [21] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, *ACM Computing Surveys* 51 (2018) 1–42. doi:10.1145/3236009.
- [22] 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. URL: <https://doi.org/10.3233/IA-210120>. doi:10.3233/IA-210120.
- [23] R. Calegari, F. Sabbatini, The PSyKE technology for trustworthy artificial intelligence 13796 (2023) 3–16. URL: [https://doi.org/10.1007/978-3-031-27181-6\\_1](https://doi.org/10.1007/978-3-031-27181-6_1). 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.
- [24] F. Sabbatini, C. Sirocchi, R. Calegari, Symbolic knowledge comparison: Metrics and methodologies for multi-agent systems, in: M. Alderighi, M. Baldoni, C. Baroglio, R. Micalizio, S. Tedeschi (Eds.), *Proceedings of the 25th Workshop “From Objects to Agents”, Bard (Aosta)*, Italy, July 8–10, 2024, volume 3735 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 202–216. URL: [https://ceur-ws.org/Vol-3735/paper\\_17.pdf](https://ceur-ws.org/Vol-3735/paper_17.pdf).



- [25] 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. 22nd Workshop “From Objects to Agents” (WOA 2021), Bologna, Italy, 1–3 September 2021. Proceedings.
- [26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research (JMLR)* 12 (2011) 2825–2830. URL: <https://dl.acm.org/doi/10.5555/1953048.2078195>.
- [27] F. Sabbatini, R. Calegari, Achieving complete coverage with hypercube-based symbolic knowledge-extraction techniques, in: S. Nowaczyk, P. Biecek, N. C. Chung, M. Vallati, P. Skruch, J. Jaworek-Korjakowska, S. Parkinson, A. Nikitas, M. Atzmüller, T. Kliegr, et al. (Eds.), Artificial Intelligence. ECAI 2023 International Workshops – XAI<sup>3</sup>, TACTIFUL, XI-ML, SEDAMI, RAAIT, AI4S, HYDRA, AI4AI, Kraków, Poland, September 30 – October 4, 2023, Proceedings, Part I, volume 1947 of *Communications in Computer and Information Science*, Springer, 2023, pp. 179–197. URL: [https://doi.org/10.1007/978-3-031-50396-2\\_10](https://doi.org/10.1007/978-3-031-50396-2_10). doi:10.1007/978-3-031-50396-2\_10.
- [28] F. Sabbatini, G. Ciatto, A. Omicini, Semantic Web-based interoperability for intelligent agents with PSyKE, in: D. Calvaresi, A. Najjar, M. Winikoff, K. Främling (Eds.), Explainable and Transparent AI and Multi-Agent Systems, volume 13283 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 124–142. URL: [http://link.springer.com/10.1007/978-3-031-15565-9\\_8](http://link.springer.com/10.1007/978-3-031-15565-9_8). doi:10.1007/978-3-031-15565-9\_8.
- [29] F. Sabbatini, R. Calegari, On the evaluation of the symbolic knowledge extracted from black boxes, *AI and Ethics* 4 (2024) 65–74. doi:<https://doi.org/10.1007/s43681-023-00406-1>.
- [30] F. Sabbatini, R. Calegari, Symbolic knowledge-extraction evaluation metrics: The FiRe score, in: K. Gal, A. Nowé, G. J. Nalepa, R. Fairstein, R. Rădulescu (Eds.), Proceedings of the 26th European Conference on Artificial Intelligence, ECAI 2023, Kraków, Poland. September 30 – October 4, 2023, 2023. URL: <https://ebooks.iospress.nl/doi/10.3233/FAIA230496>. doi:10.3233/FAIA230496.
- [31] F. Sabbatini, G. Ciatto, A. Omicini, GridEx: An algorithm for knowledge extraction from black-box regressors, in: D. Calvaresi, A. Najjar, M. Winikoff, K. Främling (Eds.), 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.
- [32] 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>.
- [33] 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. URL: <https://doi.org/10.3233/IA-230001>. doi:10.3233/IA-230001.
- [34] F. Sabbatini, R. Calegari, Symbolic knowledge extraction from opaque machine learning predictors: GridREx & PEDRO, in: G. Kern-Isberner, G. Lakemeyer, T. Meyer (Eds.), *Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning, KR 2022, Haifa, Israel. July 31 - August 5, 2022*, 2022. URL: <https://proceedings.kr.org/2022/57/>.
- [35] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, *Classification and Regression Trees*, CRC Press, 1984.
- [36] 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. Mangani, 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 (AI\*IA 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>.
- [37] F. Sabbatini, R. Calegari, Untying black boxes with clustering-based symbolic knowledge extraction, *Intelligenza Artificiale* 18 (2024) 21–34. URL: <https://doi.org/10.3233/IA-240026>. doi:10.3233/IA-240026.
- [38] F. Sabbatini, R. Calegari, Bottom-up and top-down workflows for hypercube- and clustering-based knowledge extractors, in: D. Calvaresi, A. Najjar, A. Omicini, R. Aydogan, R. Carli, G. Ciatto, K. Främling (Eds.), *Explainable and Transparent AI and Multi-Agent Systems. Fifth International Workshop, EXTRAAMAS 2023, London, UK, May 29, 2023, Revised Selected Papers*, volume 14127 of *LNCS*, Springer Cham, Basel, Switzerland, 2023, pp. 116–129. doi:10.1007/978-3-031-40878-6\_7.
- [39] F. Sabbatini, R. Calegari, Unlocking insights and trust: The value of explainable clustering algorithms for cognitive agents, in: R. Falcone, C. Castelfranchi, A. Sapienza, F. Cantucci (Eds.), *Proceedings of the 24th Workshop “From Objects to Agents”, Roma, Italy, November 6–8, 2023*, volume 3579 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 232–245. URL: <https://ceur-ws.org/Vol-3579/paper18.pdf>.
- [40] F. Sabbatini, R. Calegari, Explainable clustering with CREAM, in: P. Marquis, C. S. Tran, G. Kern-Isberner (Eds.), *20th International Conference on Principles of Knowledge Representation and Reasoning (KR 2023)*, IJCAI Organization, Rhodes, Greece, 2023, pp. 593–603. doi:10.24963/kr.2023/58.
- [41] F. Sabbatini, R. Calegari, ExACT explainable clustering: Unravelling the intricacies of cluster formation, in: C. K. Baker, L. Gómez Álvarez, J. Heyninck, T. Meyer, R. Peñaloza, S. Vesic (Eds.), *Joint Proceedings of the 2nd Workshop on Knowledge Diversity and the 2nd Workshop on Cognitive Aspects of Knowledge Representation co-located with 20th International Conference on Principles of Knowledge Representation and Reasoning (KR 2023)*, Rhodes, Greece, September 3–4, 2023, volume 3548 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: <https://ceur-ws.org/Vol-3548/paper3.pdf>.
- [42] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, in: L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu, G. Williams (Eds.), *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.