Neural-Symbolic Integration for Fairness in AI

Benedikt Wagner, Artur d'Avila Garcez

City, University of London, Northampton Square, London, EC1V0HB, United Kingdom

Abstract

Deep learning has achieved state-of-the-art results in various application domains ranging from image recognition to language translation and game playing. However, it is now generally accepted that deep learning alone has not been able to satisfy the requirement of fairness and, ultimately, trust in Artificial Intelligence (AI). In this paper, we propose an interactive neural-symbolic approach for fairness in AI based on the Logic Tensor Network (LTN) framework. We show that the extraction of symbolic knowledge from LTN-based deep networks combined with fairness constraints offer a general method for instilling fairness into deep networks via continual learning. Explainable AI approaches which otherwise could identify but not fix fairness issues are shown to be enriched with an ability to improve fairness results. Experimental results on three real-world data sets used to predict income, credit risk and recidivism in financial applications show that our approach can satisfy fairness metrics while maintaining state-of-the-art classification performance.

Keywords

Neurosymbolic AI, Deep Learning with Knowledge Representation, Fairness, Explainability

1. Introduction

Machine Learning models based on deep neural networks have demonstrated powerful classification, prediction and optimisation capability. However, they lack transparency and comprehensibility due to their large size and distributed nature. One of deep learning's main characteristics is the ability to make predictions by learning an arbitrary function in end-to-end fashion with the optimisation based solely on an input-output mapping, and with the model having flexibility to adapt all the parameters. A well-known weakness of such complex distributed representation is that they are *black boxes* when it comes to providing explanations to the predictions that were made; the set of parameter values does not offer any insight into an inherent logic of the learned model.

In response to initiatives from regulatory authorities such as the EU's General Data Protection Regulation 2016/679 [1] and societal demands for trustworthy AI and autonomous systems adhering to ethical principles, efforts have been on the increase towards achieving interpretable Machine Learning (ML) and explainable Artificial Intelligence (XAI). As ML applications are becoming increasingly influential with societal impact, fair decision-making is becoming increasingly essential for ML research. A plethora of recent papers have produced a large number

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021) - Stanford University, Palo Alto, California, USA, March 22-24, 2021.

Benedikt.Wagner@city.ac.uk (B. Wagner); A.Garcez@city.ac.uk (A.d. Garcez)

0000-0001-7375-9518 (A.d. Garcez)

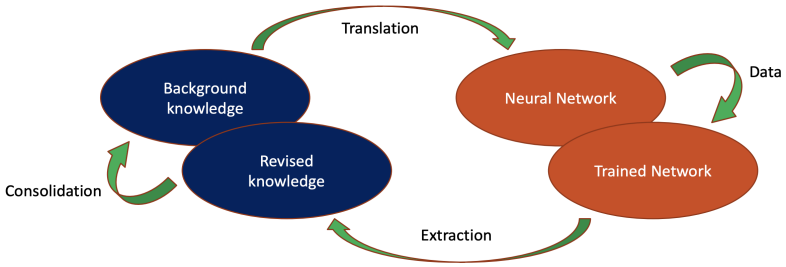

Figure 1: Illustration of the neural-symbolic cycle [10]: knowledge extraction will be carried out by querying a deep network interactively and learning continually, thus applying the neural-symbolic cycle multiple times. Knowledge inserted into the training of the deep network in the form of first-order logic constraints will be shown to improve fairness while maintaining performance of the system in direct comparison with current standard methods on different fairness metrics. The neural-symbolic cycle can be seen as offering common ground for communication and system interaction, allowing for a human-in-the-loop approach. Symbolic knowledge representations extracted from the learning system at an adequate abstract level for communication with users should allow knowledge consolidation and targeted revision, and will be evaluated in the context of improving fairness constraints.

of XAI approaches with varied results, measured in different ways, towards achieving a better understanding of the behaviour of black box AI and ML systems, often aimed at discovering undesired model properties such as unfair treatment based on protected attributes. While some approaches attempt to extract global decision-making information from a complex model (known as the *teacher* model) by learning a simpler (*student*) model - starting with the seminal TREPAN approach [2] -, others have sought to provide explanations, which have become known as local explanations, by describing specific representative cases (e.g. [3, 4]). Despite the increasing adoption of such XAI methods known as *distillation* [5], the connection with symbolic AI and knowledge representation has been largely ignored recently. In this paper, by taking a hybrid (neurosymbolic) approach, we seek to maintain a correspondence between the deep network and its counterpart symbolic description. Through the use of a neural-symbolic approach known as Logic Tensor Networks (LTN) [6, 7, 8], we can achieve interactive explainability by querying the deep network for symbolic knowledge in the form of first-order logic rules. At the same time, we can offer an approach for achieving fairness of an end-to-end learning system by intervening at the symbolic counterpart of the system with the addition of fairness constraints. With the use of the neural-symbolic cycle (Fig.2), we seek to bridge low-level information processing such as perception and pattern recognition with reasoning and explanation at a higher-level of abstract knowledge [9]. The framework introduced in this paper seeks to improve fairness via continual learning with symbolical LTN constraints.

The contributions of this paper are:

(1) We introduce a method that allows one to act on information extracted by any XAI approach in order to prevent the learning model from learning unwanted behaviour or bias discovered by the XAI approach. We demonstrate how our method can leverage an existing XAI method, known as SHAP [11], to discover and address undesired model behaviour.

(2) We implement and outline the use of LTN for continual learning and iterative querying by caching the learned representations and by using network querying in first-order logic to

check for knowledge learned by the deep neural network.

(3) We apply the proposed method and tool to the field of quantitative fairness in finance. Experimental results reported in this paper show comparable or improved accuracy across three data sets while achieving fairness based on two fairness metrics, including a 7.1% average increase in accuracy in comparison with a state-of-the-art neural network-based approach [12].

In Section 2, we position the paper in the context of the related work. In Section 3, we introduce the interactive continual-learning LTN method. In Section 4, we present and discuss the experimental results achieving fairness. In Section 5, we conclude the paper and discuss directions for future work.

2. Related Work

Explainability: The goals of achieving comprehensible Machine Learning systems are diverse [13]. There have been many recent proposals to tackle the problem differently, focusing either on the system before training or during training to obtain inherently interpretable models or post-training analyses and knowledge extraction. We shall focus on the latter two as they have gained the most attention recently and connect to our approach more closely. The most common way to differentiate XAI methods currently in practice utilises two categories: global and local explanations [14]. Most of the approaches that seek to incorporate an inherent level of interpretability are global[1]. Whereas inherently interpretable models come with stringent architectural constraints on the model itself, our approach is model-agnostic since LTN as a framework simply requires the ability to query any deep network (or any ML model) for its behaviour, that is, observing the value of an output given a predefined input, thereupon used as part of a constraint-based regularisation [7]. The predictive model itself can be chosen independently, with the LTN acting as an interface.

Post-training methods seek to achieve explanation by approximating the behaviour of complex black-box systems. One of the most prominent methods at present, based on this idea, is called LIME, which stands for Local Interpretable Model-Agnostic Explanation [4]. As the name suggests, these explanations describe specific instances by approximating local variations in the neighbourhood of a prediction. Although LIME can give very intuitive insights into predictions, it remains unclear how widely applicable local explanations are, how problematic the assumption of linearity is, and what may constitute a valid definition of soundness and measure of closeness. The reader is referred to [15] for a critique of LIME.

A method that has gained traction recently in finance is the Shapley value approach. Initially proposed by [16], it has recently been adapted to ML models by [17]. The goal is to capture the average marginal contribution of a feature value across different possible combinations. A single Shapley value for a feature of this specific input denotes the contribution of such a feature to the prediction w.r.t. the average prediction for the data set [11]. The authors propose determining such a value by calculating the average change in the prediction by randomly adding features to the model. The Shapley value works for both classification and regression tasks. [11] contributed to a significant boost in the popularity of this method by unifying various

[1] In [14], the XAI methods that are inherently interpretable are further differentiated into rule-based, prototype-based, and others.

feature attribution approaches into one framework and publishing a user-friendly implementation. We refer the reader to [14] for an extensive survey of more methods.

While the above XAI methods make a noticeable contribution to the obstacle of explainability, none of them address the problem of how one should act upon the extracted information. Consequently, we do not see the LTN approach as a method to be compared directly with the above XAI approaches but to complement them. By applying the neural-symbolic cycle multiple times, partial symbolic descriptions of the knowledge encoded in the deep network will be checked and, through a human-in-the-loop approach, incorporated into the cycle as a constraint on the learning process. This will enable an interactive integration of a desired behaviour, notably fairness constraints, by checking and incorporating knowledge at each cycle, instead of (global or local) XAI serving only to produce a one-off description of a static system.
Fairness: One of the main goals of the recent advancements in explainability encompasses considerations of the fairness of automated classification systems. Although the discovery of such unwanted behaviour is essential and useful, in this paper, we can address specific undesired properties, discovered or specified symbolically, and alter the learned model towards a fairer description.
There have been a few methods addressing fair representation or classification: [18] seek to achieve fair representation. [19] seek to achieve fair classification by proposing a reductionist approach that translates the problem onto a sequence of cost-sensitive classification tasks [19]. [20] study fairness in naive Bayes classifiers and propose an interactive method for discovering and eliminating discrimination patterns. [12] integrate fairness into neural networks by including complex and non-decomposable loss functions into the optimisation. Fairness remains a significant challenge for Machine Learning. For an overview of the variety of fairness notions, we refer the reader to [21, 18]. For an extensive overview of various fairness-oriented ML methods, we refer the reader to [22] and [23].

The above methods are related because they introduce constraints either on the data or the model during learning. The LTN-based approach used here introduces constraints as a regularisation which therefore may apply to any model or data set. Also, in LTN, additional fairness axioms can be specified during training time by the user, which may be unrelated to the existing fairness axioms. Finally, at test time, the protected variables defined by such axioms are not used, so that a final customer of the ML system will not be asked for sensitive information on gender, race, etc.

3. Method

The framework used in this paper is that of Logic Tensor Networks [6] and [8]. However, instead of treating the learning of the parameters from data and knowledge as a single process, we emphasise the dynamic and flexible nature of the process of training from data, querying the trained model for knowledge, and adding knowledge in the form of constraints for further training, as part of a cycle whose stopping criteria are governed by a fairness metric. Furthermore, we focus on the core of the LTN approach: constraint-based learning from data and first-order logic knowledge, and we make it iterative by saving the learned parametrisation at each cycle in our implementation, while removing unnecessary constraints such as the use

with LTN of Neural Tensor Networks. In our experiments, we use standard feedforward neural networks.

3.1. Language

Logic Tensor Networks [6, 24, 8] implement a many-valued first-order logic (FOL) language $\mathcal{L}$, which consists of a set of constants $\mathcal{C}$, variables $\mathcal{X}$, function symbols $\mathcal{F}$ and predicate symbols $\mathcal{P}$. Logical formulas in $\mathcal{L}$ allow to specify background knowledge related to the task at hand. The syntax in LTN is that of FOL, with formulas consisting of predicate symbols and the connectives for negations ($\neg$), conjunction, disjunction and implication ($\wedge, \vee, \longrightarrow$) and quantifiers ($\forall, \exists$).

3.2. Grounding

As for the semantics of $\mathcal{L}$, LTN deviates from the standard abstract semantics of FOL and proposes a concrete semantics where domains are interpreted in the Real field $\mathbb{R}$ as defined in *Real Logic* [6]. To emphasize that symbols are interpreted according to their grounding onto real numbers, LTN uses the term *grounding*, denoted by $\mathcal{G}$, in place of interpretation. Every object denoted by a constant, variable or term is grounded onto a tensor of real numbers. Function symbols are grounded as functions in the vector space, that is, an m-ary function maps m vectors of real numbers to one vector of real numbers. Predicates are grounded as functions that map onto the interval $[0, 1]$ representing the predicate's degree of truth.

The semantics for the connectives is defined according to fuzzy logic semantics: conjunctions are approximated by t-norms (e.g. $min(a, b)$), disjunctions by t-conorms (e.g. $max(a, b)$), negation by fuzzy negation (e.g. $1 - a$) and implication by fuzzy implications (e.g. $max(1 - a, b)$). The semantics of the quantifiers is defined by aggregation functions. For instance, in the sentence $\exists x(P(x) \wedge Q(x))$, $\exists$ can be implemented using max and $\wedge$. Krieken et al. [25] analysed various fuzzy operators and recommended those suitable for differentiable learning. In this paper, we approximate binary connectives using the product t-norm and the corresponding t-conorm and S implication. The universal quantifier is defined as the generalised mean, also referred to as p-mean[2] by Krieken et al. [25].

3.3. Learning

The objects denoted by LTN constants and variables can be learned from data. LTN functions and predicates are also learnable. Thus, the grounding of symbols depend on a set of parameters θ. With a choice of a multilayer perceptron to model each logical predicate, the parametrization

[2] $p\text{-}mean(x_1, \ldots, x_n) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}}$.

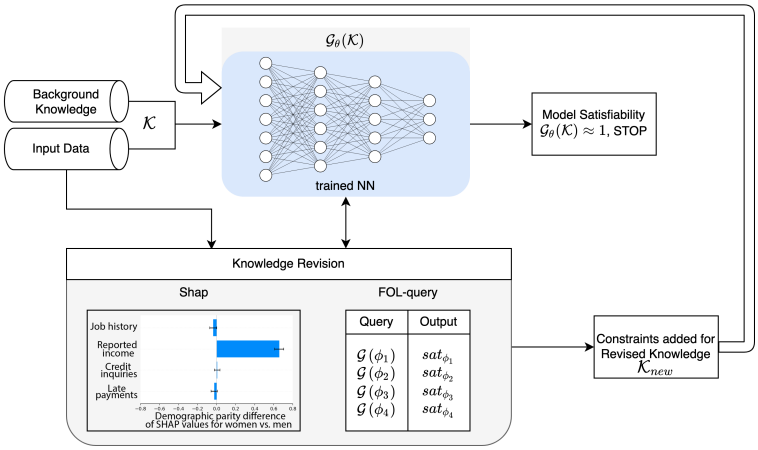

Figure 2: Illustration of the LTN pipeline for interactive continual learning: revision is carried out by querying a deep network interactively and learning continually, thus applying the neural-symbolic cycle multiple times. Explanations extracted from the network using e.g. SHAP can highlight bias in feature importance. Querying the network in LTN-style shows the satisfiability of fairness constraints which can be added to the knowledge-base K for further training. This process concludes once it has been shown to reduce bias at a subsequent SHAP explanation.

used in this paper is:

$$\mathbf{h}^{(1)}(\mathbf{v}) = g^{(1)}\left(\mathbf{v}V_P^{(1)T} + b^{(1)T}\right)$$
$$\mathbf{h}^{(2)}(\mathbf{v}) = g^{(2)}\left(\mathbf{h}^{(1)}(\mathbf{v})V_P^{(2)T} + b^{(2)T}\right)$$
$$\mathcal{G}(P)(\mathbf{v}) = \sigma\left(\mathbf{h}^{(2)}(\mathbf{v})V_P^{(3)T} + b^{(3)T}\right)$$

where each $g^{(l)}$ is an activation function, e.g. ReLU, $V^{(l)}$ is a mxn weight matrix, and $b^{(l)}$ a bias vector; σ denotes the sigmoid activation function which ensures that predicate P is mapped from $\mathbb{R}^{mxn}$ to a truth-value in $[0, 1]$.

Since the grounding of a formula $\mathcal{G}_\theta(\phi)$ denotes the degree of truth of ϕ, one natural training signal is the degree of truth of the formulas in the knowledge-base $\mathcal{K}$. The objective function is therefore to maximize the satisfiability of all formulas in $\mathcal{K}$:

$$\theta^* = \arg\max_{\theta \in \Theta} \ \mathrm{Sat}_A(\mathcal{G}_\theta(\mathcal{K}))$$

which is subject to an aggregation A of all formulas, e.g. the above-mentioned p-mean.

Notice that in the above formulation we have substituted the Neural Tensor Network used in [6] with a multilayer perceptron for simplicity. Notice also that the approach described in Figure 2 is model-agnostic. The main idea from LTN used here is that of learning with knowledge-base constraints and querying with many-valued first-order logic.

3.4. Continuous Querying

LTN inference using first-order logic clauses is not a post-hoc explanation in the traditional sense. In this paper, we argue that inference should form an integral part of an iterative process

allowing for incremental explanation through distillation of knowledge which is guided by data. We achieve this by computing the value of a grounding $\mathcal{G}_\theta(\phi_q)$, given a trained network (set of parameters θ), for a user-defined query ϕ_q.

Specifically, we save and reinstate the learned parameters stored in the LTN implementation. This is done by storing the parameters θ resulting from $\boldsymbol{\theta}^* = \arg\max_{\theta\in\Theta} \mathrm{Sat}_A\mathcal{G}_\theta(\mathcal{K})$. This also means that changes made to the knowledge-base followed by further training will not reinitialise parameters, but will instead start from saved $\boldsymbol{\theta}^*$. Having this functionality allows us to continually query and guide the learning process according to added knowledge $\mathcal{K}_{new}$, an approach akin to that of continual learning.

A query is any logical formula expressed in first-order-logic. Queries are evaluated by calculating the grounding $\mathcal{G}$ of any formula whose predicates are already grounded in the multilayer perceptron, or even by defining a predicate in terms of existing predicates. For example, the logical formula $\forall x : (A(x) \longrightarrow B(x))$ can be evaluated by applying the values of x, obtained from the data set, to the trained perceptron, obtaining the values of output neurons A and B in [0,1] (corresponding to the truth-values of predicates A and B, respectively), and calculating the implication with the use of the Reichenbach-norm and aggregating for all x using the p-mean. For an extensive analysis, we refer the reader to [25].

Algorithm 1 illustrates the steps we take to continuously refine $\mathcal{K}_{new}$ with a human-in-the-loop. The queries derive from questions a user might have about the model's response: how does the model behave for a specific group? How does the model behave for particular edge-cases? These questions can be translated relatively easily into FOL-queries. Simultaneously, an XAI method further informs the user about possible undesired model behaviour which may not be as apparent as the above common questions. This can be accomplished by a variety of XAI methods which may give insight into the functionality of a black box model. In Figure 2, XAI method SHAP reports a discrepancy in how the variable *reported income* is used by the ML system for men and women. This can be changed by adding knowledge to $\mathcal{K}_{new}$ and retraining, as will be illustrated in the experiments.

3.5. Fairness

Quantitative fairness metrics seek to introduce mathematical precision to the definition of fairness in ML. Nevertheless, fairness is rooted in ethical principles and context-dependent human value judgements. This functional dependence on value judgements is perhaps manifested in the existence of mutually incompatible definitions of fairness [26, 11]. Rather than comparing different notions of fairness, this paper focuses on achieving fairness as a desired outcome of explainability and therefore it evaluates both the classical demographic parity metric and the legal notion of disparate impact within the proposed framework of Algorithm 1.

The majority of fairness approaches in ML can be considered to target group fairness, meaning parity among groups on aggregate.[3] We adopt the following definitions of group fairness to measure and compare our approach with other methods. Following [19, 12], we consider a binary classification setting where the training examples consist of triples (X, A, Y) where $x \in X$

[3]By contrast, [18] advocate a more fine-grained individual fairness where similar individuals should be treated similarly. Although our focus in this paper is on group fairness for the sake of achieving comparative results, we believe that it should be possible to apply the approach proposed here to individual fairness.

is a feature vector, $a \in A$ is a protected attribute, and $Y \in \{0, 1\}$ is a label.

Definition 3.1. Demographic Parity (DP): A classifier h satisfies demographic parity under a distribution on (X, A, Y) if its predictions $h(X)$ are independent of the protected attribute A. That is, $\forall a \in A$:

$$\mathbf{P}[h(X) = Y \mid A = a] = \mathbf{P}[h(X) = Y]$$

Since $Y \in \{0, 1\}$, we can say that:

$$\forall a : \mathbb{E}[h(X) \mid A = a] = \mathbb{E}[h(X)]$$

The metric itself is typically reported as the difference between the above expected values which should converge to zero for a fair classifier.

Definition 3.2. Disparate Impact (DI): Given (X, A, Y) as specified above, a classifier h has disparate impact if:

$$\frac{\mathbf{P}(h(x) > 0 \mid a = 0)}{\mathbf{P}(h(x) > 0 \mid a = 1)} \leq \tau$$

Adopting the "80%-rule" from industry [27] would set the arbitrarily threshold for acceptable adverse impact to at least 80% outcome alignment. This metric compares the proportion of individuals that receive a positive output from an unprivileged and a privileged group and converges towards value 1.0 for full removal of DI between groups.

Algorithm 1: LTN-active learning cycle

Input: Data-set, Knowledge (in the form of FOL)
Output: Model satisfiability measured as overall sat-level
for *each predicate* P *in* $\mathcal{K}$ **do**
  Initialize $\mathcal{G}_\theta(P)$ `// each P can be a multilayer perceptron or output neuron`
for *epoch* < *num-epochs* **do**
  max sat $\mathcal{G}_\theta(\mathcal{K})$ `// optimize θ to achieve max satisfiability of K`
while *Revision* **do** `// user-defined Boolean`
  for *each FOL-query* ϕ_q **do**
    Calculate $\mathcal{G}\left(\phi_q\right)$ `// query the network to obtain the truth-value of` ϕ_q
    if $\mathcal{G}\left(\phi_q\right) < t$ `// t in [0,1] is a user-defined minimum sat value`
    then
      Add ϕ_q to $\mathcal{K}_{new}$
  Apply XAI-method `// we use Shapley values`
  for *each predicate* P **do**
    Inquire $\mathcal{G}_{\boldsymbol{\theta}}(P)$ `// query predicate-specific groundings`
    if $\mathcal{G}_{\boldsymbol{\theta}}$ *has undesired property* $f(\mathcal{G}_{\boldsymbol{\theta}})$ **then** `// user-determined desiderata`
      Revise $f(\mathcal{G})$ to $\mathcal{K}_{new}$ `// method-dependent revision`
  if $\mathcal{K}_{new} \neq \emptyset$ **then**
    $\mathcal{K} \leftarrow \mathcal{K} \cup \mathcal{K}_{new}$
    $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \Theta} \mathcal{G}_\theta(\mathcal{K})$ `// re-train the network`

4. Experimental Results

The following experiments illustrate how LTN can be used to obtain insight into a neural network model and interactively address an undesired behaviour by adding new knowledge to the background knowledge as illustrated in Figure 2. Background knowledge is used to provide a meaningful semantics to the explanations, facilitating human-machine interaction, while being injected into the neural network to achieve a desired property [28]. Experiment 1 is presented in a simulated environment to observe if one can achieve a desired behaviour in an idealised world. Experiment 2 provides a practical translation of the idea onto real data and reports comparisons with the state-of-the-art constraint-based learning methods for fairness. [4]

Experiment 1. Fairness using objective features: This experiment draws a parallel with a current state-of-the-art method in the area of explainability. We demonstrate how traditional XAI methods are able to benefit from a neural-symbolic approaches. Most importantly, we demonstrate how the LTN method can remove any undesired disparities in a model-agnostic approach when having access to objective features as proposed in [18].

We use the same example as the authors of the popular SHAP library connecting XAI and fairness [11]. They aim to dissect the model's input features to understand the disparities based on quantitative fairness metrics in the context of a credit underwriting scenario. A data generation process allows one to ensure that the labels are statistically independent of the protected class and any remaining disparities result from measurement, labelling or model errors. We generate four hypothetical causal factors scaled between [0-1] (income stability, income amount, spending restraint and consistency), which influence the observable features (job history, reported income, credit inquiries and late payments). The customer quality for securing credit is the product of all the factors that consequently determines the label as *high-customer quality* by being strong simultaneously in all factors. The observable features are subject to a bias introduced to obtain disparities in the system. This bias influences the mapping of the underlying factors to the observable features and therefore simulate an under-reporting of errors for women (the implementation contains further detailed explanation).

We compare the demographic disparity between the gender groups by calculating their Shapley values. We use such values as a popular way of gaining insight into model behaviour, although other explainability methods could have been used, and show that one can intervene in the model by adding knowledge for further training of the LTN to reduce disparities.

Since the SHAP method uses the same units as the original model output, we can decompose the model output using SHAP and calculate each feature's parity difference using their respective Shapley value. Then, by adding clauses to LTN which seek to enforce equality as a soft constraint, the neural network will be trained to reduce the difference (axioms 3-7 below). Reapplying SHAP would then give a measure of the success of the approach on parity difference. Axioms 3-7 are created based on the idea of treating similar people similarly from [18]. It is argued that finding an objective similarity or distance metric in practice can be challenging but should be possible[5].

First, we split the data into two subsets for the protected (F) and unprotected (M) group, re-

[4]The code for the experiments can be found here: http://github.com/benediktwagner/LTN_fairness.

[5]The authors further advocate making such metric public to allow for transparency. They propose the use of a normative approach to fairness as the absolute guarantee of fairness.

spectively, and create five subsets within each group, denoted $\mathcal{R}_{Fi}$ and $\mathcal{R}_{Mi}$, $1 \leq i \leq 5$, using quantile-based discretisation of customer quality. [6]. The five axioms (3 to 7) then state, according to the discretisation, that if a member (x) of set $\mathcal{R}_{Fi}$ defaults on credit, i.e. $h(x) = 1$, then a member (y) of set $\mathcal{R}_{Mi}$ should also default, $h(y) = 1$, and vice-versa. Given the different groups, one may wish to specify that equality in prediction is required, e.g. for the bottom 20% of the protected group w.r.t. the unprotected group according to a fairness measure. In our approach, the use of the generalised p-mean lends itself very well to this task by allowing for different forms of aggregation for each equality sub-group (referred to as *customer quality* 1 to 5 below). As a result, the user can specify in the system how strictly each fairness axiom is expected to be satisfied (using the p-mean parameter p and the satisfiability threshold t, c.f. Algorithm 1). Experiment 1 is summarised below.

Predicate: D for the positive class (i.e. credit default)

Training data: $\mathcal{T}_D$, a set of individuals who credit default; $\mathcal{T}_N$, a set of individuals who do not credit default; $\mathcal{R}_{F1}, ..., \mathcal{R}_{F5} \subset \{\mathcal{T}_D \cup \mathcal{T}_N\}$, a set of female individuals with customer quality 1 to 5; $\mathcal{R}_{M1}, ..., \mathcal{R}_{M5} \subset \{\mathcal{T}_D \cup \mathcal{T}_N\}$, a set of male individuals with the same customer quality.

Axioms:

$$\forall x \in \mathcal{T}_D : \quad D(x) \tag{1}$$

$$\forall x \in \mathcal{T}_N : \quad \neg D(x) \tag{2}$$

$$\forall x \in \mathcal{R}_{F1}, y \in \mathcal{R}_{M1} : \quad D(x) \leftrightarrow D(y) \tag{3}$$

$$\forall x \in \mathcal{R}_{F2}, y \in \mathcal{R}_{M2} : \quad D(x) \leftrightarrow D(y) \tag{4}$$

$$\forall x \in \mathcal{R}_{F3}, y \in \mathcal{R}_{M3} : \quad D(x) \leftrightarrow D(y) \tag{5}$$

$$\forall x \in \mathcal{R}_{F4}, y \in \mathcal{R}_{M4} : \quad D(x) \leftrightarrow D(y) \tag{6}$$

$$\forall x \in \mathcal{R}_{F5}, y \in \mathcal{R}_{M5} : \quad D(x) \leftrightarrow D(y) \tag{7}$$

Model: $h(x)$ (denoting $D(x)$) is the multilayer perceptron described in Section 3.3.

We initially train the multilayer perceptron on the data alone (without axioms 3 to 7) and observe the disparities shown in the SHAP XAI chart shown in 3 (left). The network learned undesired disparities among gender groups as a result of under-reporting errors in the data (equivalent to using only axioms 1 and 2 with an LTN trained for 1000 epochs).

We subsequently add axioms 3 to 7 to the knowledge-base and re-train the LTN with these axioms. As shown in the SHAP XAI chart of Figure 3 (right), this decreases disparity considerably, having reduced the Disparity Impact from 0.64 to less than 0.001. This illustrates the ability of LTN to account for fairness after having observed disparities in common XAI methods by adding appropriate fairness axioms for further training. Despite its usefulness as proof-of-concept, we acknowledge that the ideal notion of similarity among sub-groups can be impracticable. Next, we continue our investigation with additional real-world data and derive a notion of similarity automatically.

[6] This choice should be attribute-independent and application specific, e.g. based on reported income from very low, low and medium, to high and very high, different groups may require different interventions, i.e. different axioms according to policy and the situation in the real-world

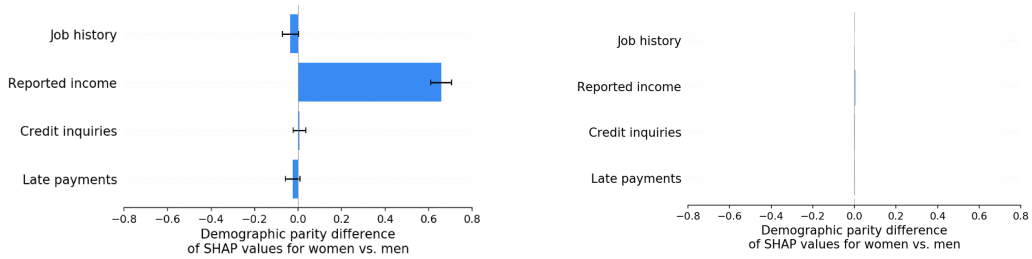

Figure 3: Disparity impact extracted with SHAP before and after LTN learning of fairness constraints.

Experiment 2. Fairness and Direct Comparisons: We compare results on three publicly-available data sets used in the evaluation of fairness, obtained from the UCI machine learning repository: the *Adult* data set for predicting income, *German* for credit risk, and *COMPASS* for recidivism. We follow the experimental setup used in [12], although they perform extensive hyper-parameter tuning whilst our models are simpler. We compare our LTN-based approach with another neural network-based approach that integrates fairness constraints into the loss function using Lagrange multipliers [12], and with an approach for naive-Bayes classifiers [20]. Gender is the protected variable in the Adult and German data sets, and race in the COMPASS data set. We train a neural network with two hidden layers of 100 and 50 neurons, respectively ([12] trains networks of up to 500 neurons per layer). We use the Adam optimiser with a learning rate of 0.001 trained for a maximum of 5000 epochs. As in [12], we report results averaged over 5-fold cross-validation.

As before, we first train the network without fairness constraints, while before we relied on an objective notion of similarity that made it possible to split the individuals into sub-groups, now we use a continual learning approach (when such objective notion is not present). The trained network is queried to return the truth-value of the predicate used for the classification task $\mathcal{G}(D(\mathcal{T}))$ for the entire training set $\mathcal{T}$. The output helps determine, as a proxy for similarity, the fairness constraints. As done in the previous experiment, a quantile-based discretisation is carried out, but this time according to the result of querying the network, after splitting the data into two subsets for each class according to the protected variable. Therefore, we obtain equally-sized groups for each protected and unprotected variable. Again, five sub-groups are used and the axioms from Experiment 1 apply.

Querying axioms 3 to 7 reveals a low sat level at first as an indication of an unfair model ($sat_{\phi_n} < 0.5$). This is confirmed by measuring the fairness metrics with DI ≤ 0.4 and DP ≥ 0.03 across all data sets [7]. The results are shown in Figure 4 which also includes the results of the approach to account for fairness in naive Bayes classifiers [20]. The comparison with [20] is not as straightforward as the comparison with FNNC [12] because [20] use 10-fold cross-validation and do not measure DI or DP. Nevertheless, we include the results of our approach using 10-fold cross-validation also in Figure 4 for the data sets. Since both LTN and FNNC are based on neural networks, we can make a more direct comparison with FNNC [12].

[7] this is also revealed in SHAP value disparity. We measure the fairness here using pre-defined metrics as fairness is a known issue in these benchmarking datasets and therefore does not require an XAI method for detection

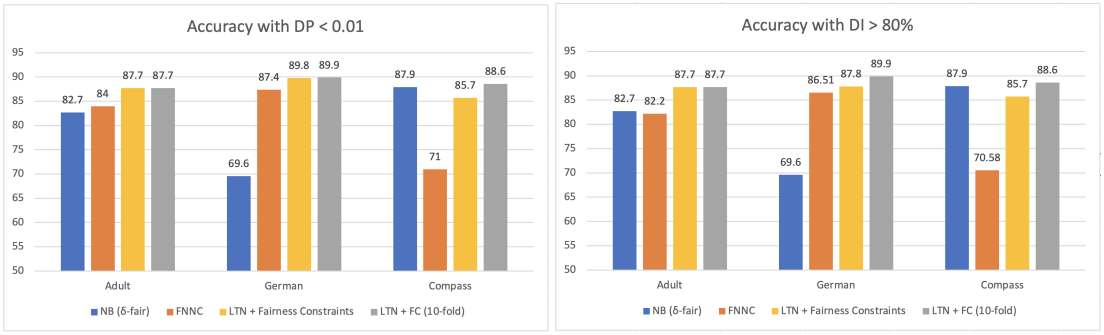

Figure 4: Comparative results of fairness-constrained learning: FNNC [12], NB-based [20] and the LTN-based approach proposed in this paper using 5-fold cross-validation (LTN + Fairness Constraints or FC) and 10-fold cross-validation (LTN + FC 10-fold), on three data sets: Adult, German and Compass. The fairness metrics only apply to FNNC and LTN; NB-based uses a different metric for fairness and is therefore not directly comparable.

As illustrated in Figure 4, we are able to outperform other state-of-the-art methods and achieve a lower variability across all data sets, and pass the DI and DP fairness thresholds proposed by [12]. All experiments were carried out using the same hyper-parameters as reported above and aggregation parameter $p = 5$. Finally, we would like to emphasize the flexibility of our approach w.r.t. different notions of fairness and its potential use with alternative fairness constraint constructions. The approach is not applicable exclusively to the metrics used here. With the increasing number and complexity of equality groups with larger p-values for aggregation, and the currently-evolving many notions of fairness being developed, we argue that rich languages such as FOL will be needed to capture more fine-grained notions, possibly converging towards individual fairness (with the generalised mean converging towards the *min* value).

5. Conclusion & Future Work

Combining XAI methods with neural-symbolic approaches allows us to not only learn about the undesired behaviour of a model but also intervene to address discrepancies which is ultimately the goal of Explainability. We have proposed an interactive model-agnostic method and algorithm for fairness and have shown how one can remove demographic disparities from trained neural networks by using a continual learning LTN-based framework. While experiment 1 demonstrated the effectiveness on addressing undesired gender-based disparities on simulated data, we have investigated such effectiveness on real-world data in experiment 2 and compared to other methods.

In the future, we plan to investigate the suitability of different XAI methods. In particular, knowledge extraction methods that extract rules directly by querying from general to specific knowledge, following the FOL language of LTN, could prove to be highly useful in practice. Furthermore, we aim to adapt this approach to varying notions of fairness as the proposed method itself is adaptable and there are lively discussions about the appropriateness of varying definitions.

Acknowledgments

We are thankful to Samy Badreddine, Michael Spranger and Luciano Serafini for their comments and useful discussions.

References

[1] European Union, Regulation 2016/679 of the European parliament and the Council of the European Union, Official Journal of the European Communities (2016).

[2] M. W. Craven, Extracting Comprehensible Models From Trained Neural Networks, Ph.D. thesis, University of Wisconsin, Madison, 1996.

[3] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. Sayres, Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV), in: 35th International Conference on Machine Learning, ICML 2018, 2018.

[4] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, arXiv:1602.04938 (2016). URL: http://arxiv.org/abs/1602.04938.

[5] N. Frosst, G. Hinton, Distilling a Neural Network Into a Soft Decision Tree (2017). URL: http://arxiv.org/abs/1711.09784.

[6] L. Serafini, A. d. Garcez, Logic tensor networks: Deep learning and logical reasoning from data and knowledge, arXiv preprint arXiv:1606.04422 (2016).

[7] M. Diligenti, M. Gori, C. Saccà, Semantic-based regularization for learning and inference, Artificial Intelligence (2017). doi:10.1016/j.artint.2015.08.011.

[8] S. Badreddine, A. d. Garcez, L. Serafini, M. Spranger, Logic Tensor Networks (2020). URL: http://arxiv.org/abs/2012.13635.

[9] A. D. Garcez, T. R. Besold, L. De Raedt, P. Foldiak, P. Hitzler, T. Icard, K. U. Kiihnberger, L. C. Lamb, R. Miikkulainen, D. L. Silver, Neural-symbolic learning and reasoning: Contributions and challenges, in: AAAI Spring Symposium - Technical Report, 2015.

[10] A. S. D'Avila Garcez, K. Broda, D. M. Gabbay, Symbolic knowledge extraction from trained neural networks: A sound approach, Artificial Intelligence (2001). doi:10.1016/S0004-3702(00)00077-1.

[11] S. M. Lundberg, S. I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, 2017.

[12] M. Padala, S. Gujar, FNNC: Achieving Fairness through Neural Networks, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, {IJCAI-20}, International Joint Conferences on Artificial Intelligence Organization, 2020.

[13] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI), IEEE Access 6 (2018) 52138–52160. doi:10.1109/ACCESS.2018.2870052.

[14] R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, F. Giannotti, A Survey Of Methods For Explaining Black Box Models, ACM Computing Surveys (2018) 1–45. URL: https://doi.org/10.1145/3236009.

[15] A. White, A. d. Garcez, Measurable Counterfactual Local Explanations for Any Classifier,

in: 24th European Conference on Artificial Intelligence, 2020. URL: http://arxiv.org/abs/1908.03020.

[16] L. S. Shapley, Contributions to the Theory of Games, in: Annals of Mathematical Studies v. 28, 1953.

[17] E. Štrumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions, Knowledge and Information Systems (2014). doi:10.1007/s10115-013-0679-x.

[18] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: ITCS 2012 - Innovations in Theoretical Computer Science Conference, 2012. doi:10.1145/2090236.2090255.

[19] A. Agarwal, A. Beygelzimer, M. Dudfk, J. Langford, W. Hanna, A reductions approach to fair classification, in: 35th International Conference on Machine Learning, ICML 2018, 2018.

[20] Y. Choi, G. Farnadi, B. Babaki, G. V. d. Broeck, Learning Fair Naive Bayes Classifiers by Discovering and Eliminating Discrimination Patterns, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019. URL: http://arxiv.org/abs/.

[21] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning fair representations, in: 30th International Conference on Machine Learning, ICML 2013, 2013.

[22] S. A. Friedler, S. Choudhary, C. Scheidegger, E. P. Hamilton, S. Venkatasubramanian, D. Roth, A comparative study of fairness-enhancing interventions in machine learning, in: FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, 2019. doi:10.1145/3287560.3287589.

[23] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A Survey on Bias and Fairness in Machine Learning (2019). URL: http://arxiv.org/abs/.

[24] I. Donadello, L. Serafini, A. D'Avila Garcez, Logic tensor networks for semantic image interpretation, in: IJCAI International Joint Conference on Artificial Intelligence, 2017. doi:10.24963/ijcai.2017/221.

[25] E. v. Krieken, E. Acar, F. v. Harmelen, Analyzing Differentiable Fuzzy Logic Operators (2020). URL: https://arxiv.org/abs/2002.06100v1.

[26] J. Kleinberg, J. Ludwig, S. Mullainathan, A. Rambachan, Algorithmic Fairness, AEA Papers and Proceedings (2018). doi:10.1257/pandp.20181018.

[27] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015. doi:10.1145/2783258.2783311.

[28] A. S. D. Garcez, L. C. Lamb, D. M. Gabbay, Neural-symbolic cognitive reasoning, Springer Science & Business Media, 2008.