Journal Title
XX(X):2–28

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

A Neurosymbolic Approach to Counterfactual Fairness

Xenia Heilmann[1,*], Chiara Manganini[2,*], Mattia Cerrato[1], Leonhard Kestel[3] and Vaishak Belle[4]

Abstract

Integrating fairness into machine learning models has been an important consideration for the last decade. Here, neurosymbolic models offer a valuable opportunity, as they allow the specification of symbolic, logical constraints that are often guaranteed to be satisfied. However, research on neurosymbolic applications to algorithmic fairness is still in an early stage. With our work, we bridge this gap by integrating counterfactual fairness into the neurosymbolic framework of Logic Tensor Networks (LTN). We use LTN to express accuracy and counterfactual fairness constraints in first-order logic and employ them to achieve desirable levels of both performance and fairness at training time. Our approach is agnostic to the underlying causal model and data generation technique; as such, it may be easily integrated into existing pipelines that generate and extract counterfactual examples. We show, through concrete examples on three benchmark datasets, that logical reasoning about counterfactual fairness has some important advantages, among which its intrinsic interpretability, and its flexibility in handling subgroup fairness. Compared to three recent methodologies in counterfactual fairness, our experiments show that a neurosymbolic, LTN-based approach attains better levels of counterfactual fairness.

Keywords

Counterfactual Fairness, Neurosymbolic AI, Knowledge Extraction

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

1 Introduction

In the last decade, there has been a considerable amount of research on the topic of fairness in deep learning, as neural networks are increasingly used in critical contexts such as credit scoring, risk assessment of recidivism, and job recruitment. As of today, making these systems fairer is a complex and multi-faceted challenge.

Considering fairness and bias in machine learning, there are two major aspects of interest: how to detect bias and how to mitigate it. A challenge regarding bias detection is that the many fairness criteria explored are sometimes incompatible (Verma and Rubin 2018; Castelnovo et al. 2021). To a large extent, the fairness area has divided itself into the two broad categories of group-based and individual-based notions, where the former see *groups of individuals* – rather than *single individuals* – as the ultimate objects of unfairness. A common objection to group-based fairness metrics is that they secure fairness for "the average individual" of a sensitive group, at the cost of ignoring existing unfairness *within* such groups, which contrasts with the main intuition that fairness has to do with treating similar individuals in a similar way (Dwork et al. 2012). To address these *desiderata*, some individual-based measures of fairness have been proposed. Among them, counterfactual fairness (CF) (Kusner et al. 2017) reframes the problem of algorithmic fairness through the lenses of causality, namely, as the counterfactual question: "Would I be treated in the same way, had my protected feature been different?".

The question regarding bias mitigation concerns the level at which we want to mitigate bias, if before training (pre-processing), at training time (in-processing), or after (post-processing) (Hort et al. 2022; Caton and Haas 2024). Pre-processing techniques comprise different transformations of the training data towards a more balanced dataset. The rationale behind it is, that a model that is trained on fair data will deliver fair predictions (Kusner et al. 2017). Post-processing handles bias of a trained model by correcting its input, the model itself, or its output (Hort et al. 2022). As they rather aim for (hard) correction than (soft) mitigation, these techniques are useful for constraints that must hold universally. In-processing comprises methods like regularization, adversarial learning, model composition, and adjusted learning methods (Hort et al. 2022; Caton and Haas 2024). Instead of simulating a desired world or correcting biased predictions, it aims

[1]Institute of Computer Science, Johannes Gutenberg University, Mainz, Germany
[2]Department of Philosophy, University of Milan, Milan, Italy
[3]Ludwig-Maximilians-Universität München, Germany and Munich Center for Machine Learning (MCML)
[4]School of Informatics, University of Edinburgh, Edinburgh and Alan Turing Institute, London, United Kingdom
[*]These authors contributed equally.

Corresponding author:
Xenia Heilmann
Email: xenia.heilmann@uni-mainz.de

to induce intrinsically fair models that are able to handle unfair data (Wan et al. 2023). A practical benefit is that in-processing techniques can also be applied to pre-trained models for (fairness) constraint learning (Wan et al. 2023).

In this context, the idea of leveraging neurosymbolic approaches to tackle algorithmic unfairness has been largely underexplored so far. The potential for a good fit between these two research lines has been pointed out in recent surveys by Gibaut et al. (2023) and Bhuyan et al. (2024). Neurosymbolic AI allows one to reason symbolically about the neural network's behaviour, by establishing a correspondence between its low-level information processing and high-level logical reasoning (Hitzler and Sarker 2022; Sarker et al. 2021). As such, the approach shows many advantages for establishing trust in deep learning systems, by making models more interpretable and transparent (Gibaut et al. 2023). Furthermore, most in-processing bias mitigation frameworks adjust the loss function or the learning algorithm of machine learning models according to a distinct, hard-coded, notion of fairness. Instead, neurosymbolic models offer flexibility, as they provide an interface between arbitrary formalised constraints and their implementation into the machine learning process.

To bridge this gap, we propose an in-processing method to train a counterfactually fair neural network by means of the neurosymbolic method of Logic Tensor Networks (LTN) proposed by Badreddine et al. (2022). Specifically, we integrate counterfactual fairness into the neural network learning process, in the form of logical constraints. Furthermore, we show how to exploit symbolic reasoning after network training to better secure fairness for specific sensitive subgroups. Lastly, we integrate a counterfactual knowledge extraction method into the LTN training process. We investigate how counterfactual explanations may be employed to reason about which features, *had they been different*, result in different outcomes, and show how to extract constraints to improve counterfactual fairness from this. We evaluate our method on three benchmark datasets with binary as well as score-based predictions. We find that our method is able to take accurate decisions with minimal infractions in terms of counterfactual fairness, especially when subgroup fairness is considered.

The main contributions of this work are the following. First, we push the state-of-the-art in neurosymbolic fairness approaches by showing how to integrate counterfactual fairness and subgroup counterfactual fairness into LTN. Secondly, we introduce a novel methodology to automatically extract fairness constraints from counterfactual explanations. Finally, we show how LTN may be employed to provide individuals insights on their outcome. A short version of this paper was previously published at the NeSy conference 2025 (Heilmann et al. 2025).

2 Preliminaries

Our pipeline is based on two main algorithmic parts: LTN to specify counterfactual fairness constraints, and counterfactual knowledge extraction based on counterfactual explanations, in the dual role of injecting further fairness constraints and inspecting the impact thereof. Here, we provide a preliminary discussion on counterfactual fairness, LTN, and counterfactual explanations.

2.1 Counterfactual Fairness

Kusner et al. (2017) introduced the notion of counterfactual fairness, according to which a classifier treats individuals fairly if they would have received the same outcome, had their sensitive attribute been different. Such a counterfactual outcome requires knowledge of the causal model $\mathcal{M}$ underlying the data-generating process. A causal model is a pair $(\mathcal{S}, \mathcal{F})$ defined as follows. Signature $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$ where $\mathcal{U}$ is the set of *exogenous* variables, i.e., variables that causally depend on factors that are outside the scope of the model such as noise and background conditions; $\mathcal{V}$ is the set of *endogenous* variables, whose values are causally determined by other variables of the model; $\mathcal{R}$ is a function that associates each variable $X \in \mathcal{U} \cup \mathcal{V}$ with the non-empty set of its possible values. Finally, $\mathcal{F}$ is the set of structural equations that determine the value of Y as a function of those of the other endogenous and exogenous variables, i.e., $\mathcal{F}_X : \mathcal{R}(\mathcal{U} \cup \mathcal{V} - \{X\}) \rightarrow \mathcal{R}(X)$.

Let us assume that, in $\mathcal{M}$, variable $\hat{Y}$ corresponds to a model's prediction and S to a binary sensitive attribute with possible values $\mathcal{R}(S) = \{s, s'\}$. The computation of the counterfactual outcome for an individual with sensitive attribute s corresponds to the intervention $\hat{Y}_{S \leftarrow s'}$ (Pearl and Mackenzie 2018). This denotes the value of the predicted outcome $\hat{Y}$ as determined by a minimally modified version of $\mathcal{M}$ in which a new structural equation of S overrides its value to s'. The formalisation by Kusner et al. (2017) requires that the probability distribution of model predictions is the same in the actual world, where $S = s$, and in the counterfactual world, where $S = s'$. This must hold for any individual i.e., under any assignment of sensitive feature S and non-sensitive feature(s) $X = (X_i, \ldots, X_n)$ in the actual world:

$$P(\hat{Y} = y | S = s, X = x) = P(\hat{Y}_{S \leftarrow s'} = y | S = s, X = x) \tag{1}$$

2.2 Causal Normalizing Flows

In real-world settings, it is infeasible to access the complete structural causal model $\mathcal{M}$ underlying the data. Therefore, recent approaches aim to approximate unknown structural causal models and generate counterfactual data based on these approximations (Kocaoglu et al. 2018; Kim et al. 2021; Grari et al. 2023; Javaloy et al. 2023). To estimate the counterfactuals of (factual) observations, one method is Causal Normalizing Flows (CNF) (Javaloy et al. 2023). CNF are causal generative models that leverage on the deep-learning method of normalizing flows to accurately and efficiently approximate $\mathcal{M}$ of a data-generating process. The approximation is carried out on the basis of (factual) observations and the causal graph induced by $\mathcal{M}$. The causal graph of $\mathcal{M}$ is a directed acyclic graph whose nodes are labeled by the endogenous and exogenous variables of $\mathcal{M}$, and where each directed edge from node a to node b indicates that the latter *depends* on the former. The exogenous variables correspond to the roots of the graph. Unlike $\mathcal{M}$, the causal graph induced by it, is in many cases obtainable through domain knowledge, as it is a description of the causal dependencies of $\mathcal{M}$, without specifying its structural equations.

2.3 Counterfactual Explanations

The counterfactual explanation of a negatively predicted data point is generally defined as the set of minimal changes to that point sufficient to obtain a positive outcome. Since we are interested in the causality of such an outcome, we will focus only on the set of minimal changes that constitute interventions on the structural causal model $\mathcal{M}$ in the technical sense described in Section 2.1. Counterfactual explanations of this kind have also been termed *consequential recommendations* in Karimi et al. (2022).

Counterfactual explanations can concern *actionable* features, i.e., features that individuals are actually able and willing to change in order to achieve a favourable outcome (like job, education level, address, etc.) or *immutable* ones, i.e., features that either cannot be actively changed (such as race, country of birth, age, gender) or because it would morally unacceptable to ask to do so (e.g., religion, marital status). Sensitive attributes are often immutable, and the definition of counterfactual fairness given in Section 2.1 precisely requires that counterfactual explanations of the model do not involve sensitive features. However, it has been observed that only a small and often insufficient set of characteristics is treated as sensitive (Simson et al. 2024). For this reason, in 4.3 we tackle the issue of counterfactual explanations involving immutable features.

2.4 Logic Tensor Networks

Logic Tensor Networks (LTN) are a neurosymbolic framework introduced by Badreddine et al. (2022), enabling generalization and inference from data by defining e.g., a neural network's loss function using logical formulas. More specifically, LTN integrates a fully differentiable first-order logic $\mathcal{L}$ with a fuzzy semantics.

Its signature includes a set of constants $\mathcal{C}$, function symbols $\mathcal{F}$, variables $\mathcal{X}$, and predicate symbols $\mathcal{P}$. These symbols are interpreted, or grounded (denoted by $\mathcal{G}$), onto tensors of real numbers, representing the domain of discourse. Constants are grounded to tensors representing datapoints or elements of the domain, while variables are grounded as finite sequences of tensors representing possible values. Functions are grounded as computations, so mathematical functions taking and returning tensors, and predicates are grounded as functions mapping tensors to a real number in $[0, 1]$, representing the degree of truth of a statement. These predicates can e.g., be realized with neural networks.

LTN employ fuzzy semantics, where any value in $[0, 1]$ is a valid truth degree. Logical connectives and quantifiers are defined accordingly: connectives such as e.g., $\wedge$ is defined as $u \wedge v = uv$. Implications are modelled by the Reichenbach implicaton $u \rightarrow v : 1 - u + uv$ and quantifiers include the existential quantifier $\exists$ as a generalized p-mean $(\frac{1}{n}\sum_{i=1}^{n} u_i^p)^{1/p}, p \geq 1$ and the universal quantifier $\forall$ as $1 - (\frac{1}{n}\sum_{i=1}^{n}(1 - u_i)^p)^{1/p}, p \geq 1$. Here, the parameter p influences the quantifier's behaviour with $p = 1$ corresponding to the arithmetic mean, $p = 2$ to root mean square, and as p approaches infinity, maximum/minimum operations are approached. Note that the selection of connectives and quantifiers presented here aligns with the stable configuration of LTN proposed by Badreddine et al. (2022), though alternative definitions exist, and the choice of operators

impacts training stability and gradient behaviour (Badreddine et al. 2022; Wagner and d'Avila Garcez 2021).

Learning in LTN involves maximizing the truth degree of logical formulas, called *axioms*, comprising a knowledge-base $\mathcal{K}$. This is achieved by aggregating the truth degrees of individual axioms, commonly using a p-mean (like the existential quantifier) via an aggregation operator. The resulting aggregated degree is the *satisfaction* (sat) of the knowledge-base. Training LTN maximizes sat (treating it as the complement of loss) using gradient descent, optimizing $1 - sat$ instead of a traditional loss function.

3 Related Work

3.1 Fairness Through Neurosymbolic Methods

The paper by Wagner and d'Avila Garcez (2021) has recently inaugurated a line of research that combines algorithmic fairness with neurosymbolic aspects. Here, the authors propose a general method for instilling fairness constraints into deep network classifiers. They apply the LTN framework and inject these fairness constraints as logically expressed axioms. Then, the learning process feeds back until these are satisfied. Their work focuses on the group fairness metrics of demographic parity (i.e., the difference between the positive outcome rates of the disadvantaged and the advantaged group) for which the reported experiments reveal that fairness with respect to these metrics is achieved without sacrificing accuracy.

The reported experiments reveal comparable or even improved accuracy across three different sets (Adult, German, and COMPAS) while achieving demographic parity, in comparison with a state-of-the-art neural network-based approach. Closely related to Wagner and d'Avila Garcez (2021), the work by Greco et al. (2023) experimentally shows that the effectiveness of LTN for securing fairness is highly dependent on the semantic interpretations chosen, and that the optimal combination of them yields results in line with previous non-neurosymbolic approaches to group fairness. While both works focus on group-based notions of fairness, the integration of counterfactual fairness into neurosymbolic frameworks has not yet been researched to the best of our knowledge.

3.2 Approaches to Counterfactual Fairness

Among the existing approaches to counterfactual fairness (Kusner et al. 2017), the majority of the work proposes to enforce it by generating counterfactual data, and then use this data to enhance factual training data to input into a machine learning training pipeline (Javaloy et al. 2023; Zuo et al. 2023; Louizos et al. 2017; Kim et al. 2021; Lin et al. 2024; Xu et al. 2019; Kocaoglu et al. 2018; Yang et al. 2021). The main focus throughout these approaches lies on the counterfactual generation process leaving aside modifications on the final predictor itself. Differently, Grari et al. (2023) claim that additionally integrating counterfactual fairness objectives into the loss function of the machine learning pipeline contributes to counterfactually fairer predictions. Our proposal builds on the latter suggestion and develops the idea of a neurosymbolic approach in which the requirements of counterfactual fairness

are expressed logically and injected at training time. We integrate counterfactual fairness by estimating counterfactuals examples via current methods; we then develop a neurosymbolic approach that can achieve an overall counterfactual fair model. Broadly speaking, what makes a neurosymbolic approach different from these other approaches is that, in the neurosymbolic model, we express the constraints symbolically and try to ensure that the prediction guarantees the satisfaction of these constraints. Some models might try to implicitly ensure correctness with respect to the fairness criteria, but this is hard to verify. This is why a neurosymbolic approach is particularly promising for ensuring fairness criteria in machine learning models and underscores the contribution of this paper.

3.3 Integrating Counterfactual Explanations

One active line of research explores the possibility of using XAI methods to detect and even mitigate violations of fairness (Deck et al. 2024). For instance, in the already mentioned pipeline by Wagner and d'Avila Garcez (2021), the SHAP explainability method (Lundberg and Lee 2017) is used, but it plays no active role in it, as it is only employed to isolate problematic imbalances and subsequently check the efficacy of their fairness constraints in their mitigation. In contrast, in our pipeline we can exploit explainability methods for the *automatic generation and injection* of *ad hoc* fairness constraints into the network.

An important explainability method is that of counterfactual explanations, that is, the set of minimal changes to a data instance sufficient to obtain a different classification outcome. Goethals et al. (2024) have introduced a method based on counterfactual explanations to detect significant patterns of discrimination. Namely, the method compares the distribution of counterfactual explanations between sensitive groups. As an example, they show that in the Adult dataset (Becker and Kohavi 1996), women are more frequently returned *marital-status*=*"husband"* as a counterfactual explanation than men. Since it is problematic to suggest that an individual should change immutable features of this type to obtain a positive outcome, we consider undesirable all those counterfactual explanations that suggest to change an immutable feature, as we argue that it is unethical in itself to suggest individuals to change features such as marital status, or religion to obtain a favourable outcome.

For this reason, in Section 4.3, we develop a method to automatically smooth out possible imbalances in undesirable counterfactual explanations between sensitive groups. We integrate the method by Goethals et al. (2024) into our pipeline with minor adjustments: On the algorithmic level, for every negatively-predicted data point $x \in \mathcal{D}^-$, we iterate over every possible value[1] f of every feature F and calculate the new prediction of this counterfactual. Subsequently, the results are aggregated on the basis of the sensitive attribute, hence highlighting differences in the distribution of counterfactual explanations between groups.

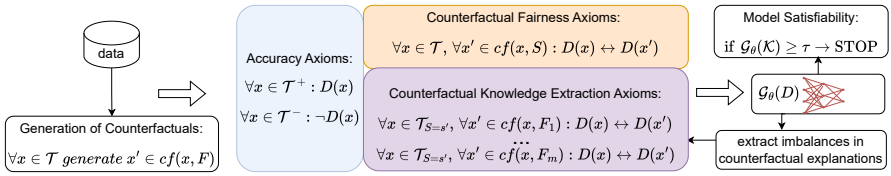


Figure 1. Overview of our pipeline for binary predictions. Here, S denotes the set of sensitive attributes, F_i a feature, $\mathcal{T}$ the dataset and $\mathcal{D}$ the prediction model.

4 Method

The goal of our pipeline is to enforce counterfactual fairness, while preserving the accuracy of predictions, and additionally disincentivising undesirable counterfactual explanations that suggest individuals to intervene on immutable features to achieve a favourable outcome. Specifically, we define certain data columns as *immutable* if they are either sensitive or particularly challenging for individuals to act upon. We achieve these three goals by integrating adequate axioms into the training process of the LTN framework. An overview of the pipeline for datasets with binary outcomes can be found in Figure 1.

As a first pre-processing step, we approximate counterfactual examples for all data points x in the data $\mathcal{T}$. Here, any counterfactual generation method can be applied, allowing flexibility in how we define counterfactuals based on the specific application and dataset. Secondly, we introduce axioms to ensure accuracy. These axioms ensure that the model retains its predictive power while incorporating fairness constraints. Following this, we introduce axioms enforcing counterfactual fairness, requiring that a datapoint and its counterfactual with respect to the sensitive attribute receive the same predicted outcome (or a similar outcome in a score-based prediction).

Finally, we optionally add axioms derived from our counterfactual knowledge extraction method, which disincentivise counterfactual explanations that recommend an intervention on an immutable feature. These axioms encode the knowledge that altering immutable attributes should not be proposed as a path to a more desirable outcome. We then train a model within the LTN framework, leveraging gradient descent to maximize the satisfaction of all axioms.

This trained model can be *post-hoc* queried for imbalances between sensitive subgroups or individual datapoints, allowing for an post-hoc analysis of potential biases. Importantly, the pipeline is iterative so that the results of this analysis can be fed back into the training pipeline by adding additional, targeted axioms, and retraining our model until a sufficient level of model satisfiability is reached. This approach allows for refinement of the fairness constraints and continuous improvement of the model's behaviour. Our pipeline is capable to handle, with different sets of axioms, both binary predictions and score-based ones. Furthermore, the axiom-based approach allows for transparent and auditable fairness interventions, enabling practitioners to understand why

the model behaves in a particular way and to tailor the fairness constraints to their ethical considerations.

4.1 Accuracy Axioms

The first axioms we add to the training pipeline ensure the accuracy of the model predictions. Here, for binary predictions, we adapt the axioms for predictive performance by Wagner and d'Avila Garcez (2021). Let D denote our classifier, $\mathcal{T}$ our dataset and let $x \in \mathcal{T}$ hold. Furthermore, let $\mathcal{T}^+$ be the set of data points with a positive outcome as ground truth and $\mathcal{T}^-$ the data points with a negative outcome as ground truth. Then, we can state the following axioms:

$$\forall x \in \mathcal{T}^+ : \quad D(x) \tag{A1}$$

$$\forall x \in \mathcal{T}^- : \neg D(x) \tag{A2}$$

Axiom A1 states that for all data points with a positive ground truth label $x \in \mathcal{T}^+$, the classifier $D(x)$ should predict a positive outcome. Conversely, Axiom A2 states that for all data points with a negative ground truth label $x \in \mathcal{T}^-$, the classifier should predict a negative outcome.

For a score-based prediction, where the output is a continuous value, our axioms have to take into account that predictions and ground truth are close. We therefore define a predicate for the equality $Eq(\hat{y}, y) = 1/(1 + 0.5 \sum_j (\hat{y}_j - y_j)^2)$, where $\hat{y}$ denotes the predicted score of the data points $x \in \mathcal{T}$ and y is the ground truth score. This predicate returns a value close to 1 when the predicted and true scores are similar, and closer to 0 as the difference increases. With this predicate defined, we have the axiom optimizing the predictive performance for score-based settings:

$$\forall x \in \mathcal{T} : Eq(D(x), y) \tag{A3}$$

4.2 Counterfactual Fairness Axioms

By adding the axioms, we want that a data point and its counterfactual with respect to the sensitive attribute S receive the same outcome. Let $x' \in cf(x, S)$ denote the set of generated counterfactuals of x with respect to the sensitive feature S. Intuitively, for counterfactual fairness to hold, the following should be true for all data points:

$$\forall x \in \mathcal{T}, \forall x' \in cf(x, S) : D(x) \leftrightarrow D(x') \tag{A4}$$

This axiom guarantees an overall counterfactually fairer model as it reformulates the original definition of counterfactual fairness expressed in Equation 1, as a first-order logic constraint. It states that for every data point x and its counterfactual x', the classifier's prediction for x must be logically equivalent to the prediction of x'. For score-based prediction, we modify this axiom to account for the continuous output. Instead of requiring the equivalence above, we check for closeness of predicted scores using the Eq predicate defined above. We reformulate $D(x) \leftrightarrow D(x')$ to $Eq(D(x), D(x'))$. This modification applies to all subsequent axioms.

We now go one step further, showing how to integrate counterfactual fairness axioms for subgroups (or "subgroup counterfactual fairness"). The rationale here is that the general Axiom A4 doesn't account for potential fairness disparities between different subgroups within the dataset. A model might enhance fairness more for one subgroup than for another, a behaviour that is not captured by a global fairness constraint. We therefore refine our axioms with respect to subgroups $C_1, \ldots, C_n$ as follows:

$$\forall x \in \mathcal{T}_{C_1}, \forall x' \in cf(x, S) : D(x) \leftrightarrow D(x') \tag{A4$_1$}$$

$$\ldots$$

$$\forall x \in \mathcal{T}_{C_n}, \forall x' \in cf(x, S) : D(x) \leftrightarrow D(x') \tag{A4$_n$}$$

This set of axioms applies the counterfactual fairness constraint separately to each subgroup. For each subgroup C_i, these axioms state that for all data points x belonging to that subgroup ($\mathcal{T}_{C_i}$) and their corresponding counterfactuals x', the classifier's predictions must be equivalent. In a simple setting, the data could be divided into subgroups based on different sensitive values. However, more refined subgroups are also supported, allowing for partitioning based on combinations of features. For example, we can further devide the sensitive groups (e.g., females and males) into subgroups based on other features (e.g., age). These subgroups can be designed to partition the entire dataset (for instance, "young females", "elderly females", "young males", and "elderly males"), or to isolate a specific subset of interest within the sensitive group, for which we want to enforce the fairness constraint. This is especially interesting in real-world scenarios where counterfactual fairness might not be relevant for all subgroups of a sensitive feature, but only for some of them. For instance, a financial institute might want to evaluate the counterfactual fairness w.r.t. gender of a loan that can be granted to young people only, or certain professionals only (e.g., teachers). In such cases, they would apply Axiom A4$_n$ for $\forall x \in \mathcal{T}_{young}$ or $\forall x \in \mathcal{T}_{teachers}$. This setup makes our approach adaptable to many applications in which subgroup counterfactual fairness is desired.

4.3 Counterfactual Knowledge Extraction Axioms

The overall idea for axioms from counterfactual knowledge extraction (CKE) is that we want to smooth out observed imbalances between sensitive groups in the frequency of certain undesirable counterfactual explanations. We consider undesirable all those counterfactual explanations $x' \in cf(x, F)$ where F is an immutable (but non-sensitive) feature, as we argue that it is unethical in itself to suggest individuals to change features such as marital status, or religion to obtain a favourable outcome. These counterfactual explanations not only raise ethical concerns but also potentially mask underlying unfairness with respect to the sensitive group. For example, Goethals et al. (2024) found that women were more frequently returned *marital-status="husband"* than men as a counterfactual explanation in the Adult dataset.

This points to a critical issue: the model may be learning to associate certain outcomes with immutable characteristics and then recommending changes to those characteristics as a pathway to a more favourable result. This not only reinforces the idea that certain

identities or attributes are inherently less desirable, but also signals underlying problems within the model itself. The very presence of these imbalances should raise red flags and prompt a thorough analysis before deployment. Such recommendations can suggest that the model has amplified biases present in the training data, or that the features are interacting in unintended ways. Our approach seeks to address this by firstly discovering such issues and secondly discouraging the generation of such problematic counterfactual explanations. The goal is not simply to achieve statistical fairness, but to ensure that the model reasons fairly and doesn't perpetuate harmful biases.

We hence want our pipeline to be able to detect an imbalance in the frequency of undesirable counterfactual explanations between sensitive groups and automatically generate *ad hoc* axioms to mitigate such an imbalance. To this end, we generate counterfactual explanations of negatively predicted data points.[2] We then compare the frequencies of counterfactual explanations across groups by aggregating the data points on the basis of the sensitive attribute, obtaining a score representing the difference of frequencies for undesirable explanations. This score provides an analyst with valuable information on which specific discrimination patterns should be addressed and for which sensitive class.

Let us denote these explanations with $(s', F_1), \ldots, (s', F_m)$ and the datapoints for which the sensitive attribute S is s' that obtained a negative prediction with $\hat{\mathcal{T}}^{-}_{S=s'}$. Then we want for a negatively predicted (factual) datapoint $x \in \hat{\mathcal{T}}^{-}_{S=s'}$ that its counterfactual explanations $x' \in cf(x, F_i)$ with respect to feature F_i, receives the same outcome as x. This indicates that feature F_i is not relevant for the outcome of the prediction. This can be modeled by the following axioms:

$$\forall x \in \hat{\mathcal{T}}^{-}_{S=s'}, x' \in cf(x,F_1) : D(x) \leftrightarrow D(x') \tag{A5$_1$}$$

$$\ldots$$

$$\forall x \in \hat{\mathcal{T}}^{-}_{S=s'}, x' \in cf(x,F_m) : D(x) \leftrightarrow D(x') \tag{A5$_m$}$$

These axioms enforce that for a given sensitive group (defined by $S = s'$), if a data point x is predicted in a certain way, then any counterfactual of x created by intervening on an immutable feature (F_i) should receive the same prediction.

While for counterfactual fairness axioms we add all axioms simultaneously in the training pipeline, these knowledge extraction-based axioms are added iteratively for better model surveillance and to oversee their individual influence to counterfactual fairness. Furthermore, a human-in-the-loop may be integrated in this part of the pipeline to assess which constraints are desirable to be integrated as axioms. It's also worth noting that explainability axioms can be applied in two ways: either in conjunction with counterfactual fairness axioms, or individually as minimal interventions to improve counterfactual fairness.

4.4 Post-Hoc Queries

Integrating counterfactual fairness into a neurosymbolic framework poses several advantages, primarily due to the framework's inherent ability to reason about logical

statements and their truth values. In particular, the satisfaction level to any logical query may be straightforwardly computed. This is of particular benefit in fairness-sensitive applications, offering a transparent way to assess and enforce fair behaviour. LTN allow us not only to ensure fairness but also to probe the model's reasoning and understand why it makes certain decisions. Here, we elaborate on two post-hoc queries.

Firstly, after training our pipeline, an individual can run an existence query to investigate potential unfairness affecting them directly. For instance, they can ask *is there a similar point in my subgroup which has a different outcome?* Concretely, for an individual data point $\hat{x}$ which is in subgroup $\mathcal{T}_C$ this query could logically formulates as:

$$\exists x \in \mathcal{T}_C, x \neq \hat{x} : \neg D(x) = D(\hat{x}) \wedge ||x - \hat{x}||_2 < \beta \tag{2}$$

Here, β denotes the parameter for *similarity*, defining how close a data point must be to $\hat{x}$ to be considered similar, and can be defined application-specific. A high satisfaction level for this query indicates the existence of (many) similar individuals within the same subgroup who received different predictions.

Secondly, the evaluation of CF can be flexibly queried for specific subgroups. This is especially interesting in applications where CF might not be relevant for all subgroups in the dataset as the application is specifically designed for one subgroup, e.g., giving out loans to teachers. Here, one can run a universally-quantified query for this subgroup and evaluate if the model is counterfactual fair with respect to the sensitive attribute. This allows for a targeted assessment of fairness, focusing on the groups most relevant to the application.

5 Research Questions

To assess our approach, we conducted experiments to showcase that integrating accuracy, CF and axioms from counterfactual knowledge extraction is beneficial for training counterfactually fairer models. The experiments address the following research questions:

(Q1) *How does our method improve counterfactual fairness, overall and at the subgroup level?* Here, we first aim to assess how well our approach reduces bias overall but also across different subgroups within the dataset.

(Q2) *How does our method compare to other approaches in terms of fairness and accuracy?* To tackle this question, we compare our pipeline against existing CF methods, evaluating its performance across both fairness metrics and predictive accuracy.

(Q3) *Can counterfactual knowledge extraction be exploited to learn effective axioms?* We explore whether automatically generated axioms based on observed imbalances in counterfactual explanations can contribute to improved fairness.

(Q4) *What can we learn from post-hoc queries?* Here, we explore the potential of the neurosymbolic framework to provide deeper insights into the model's decision-making process.

Figure 2. Partial causal graphs for the COMPAS (a) and Lawschool (b) datasets. The arrows connecting nodes and rectangles indicate that the node is connected to *every* node inside the rectangle. White nodes denote immutable sensitive features, grey nodes immutable non-sensitive features, and blue nodes actionable features.

6 Experimental Setup

In this section we introduce the datasets we tested our method on, the counterfactual generation method, the different baselines and our evaluation metrics. Our code can be found at `https://github.com/xheilmann/CounterfactualFair_LTN`.

6.1 Datasets

We conduct experiments across three benchmark datasets, as detailed below. Firstly, we ran experiments on the **Adult** dataset (Becker and Kohavi 1996), with *gender* as our sensitive attribute. This dataset contains features for binary prediction of an adult's income, comprising 48842 records with 14 features and a binary target indicating income exceeding 50K. Features include race, age, workclass, education, marital status, and occupation. As subgroups, we take each attribute combination of (*gender, race*). As immutable features we identify *marital-status, relationship, race* and *native-country*. The test set comprises around 10K data points, each accompanied by their corresponding counterfactual instances.

Furthermore, we apply our method on the **COMPAS** dataset (Angwin et al. 2016) with *race* as the sensitive attribute. Collected as part of the ProPublica analysis of machine bias in criminal sentencing, COMPAS contains 6172 observations with features relating to defendants, including age, race, prior criminal records, and a recidivism risk score. The target variable indicates rearrest within two years. Here, we evaluate on and add subgroup fairness axioms for *race* and each attribute combination of (*race, age*). For the latter, we group the *age* attribute into four categories, namely, under 30, 31-45, 46-60, and older than 60 years (**COMPAS(age)** in the following). As immutable feature we have *age*. In our test set we have on average 1230 data points and their counterfactuals.

As a third dataset, we employ the **Lawschool** dataset (Wightman 1998) with *gender* as the sensitive attribute. The Lawschool dataset consists of data on law school applicants, focusing on undergraduate GPA and LSAT scores alongside admission status. We want to stress that Kusner et al. (2017) show that this dataset is counterfactually fair with respect to gender. Here, we evaluate how adding our (subgroup) counterfactual fairness

axioms improve subgroup fairness for all combinations of (*gender, race*). We have *race* as immutable feature. We evaluate on a test set containing 4359 data points.

6.2 Counterfactuals

Our function for approximating counterfactual examples $x' \in cf(x, F)$ is implemented via causal normalizing flows (Javaloy et al. 2023) as explained in Section 2.2. We extended the original code[3] to generate the counterfactuals with respect to *all* the attributes, rather than just for the sensitive one. This is necessary for the generation of the counterfactual knowledge extraction (CKE) axioms described in Section 4.3. For the interventions, we set the feature value to the most frequent values that are at least present in 1% of the dataset, up to a maximum of 10 values. For continuous features, we took their percentiles. For the experiments, we used the partial[4] causal graphs by Zhang et al. (2016) for Adult, by Russell et al. (2017) for COMPAS, and by Kusner et al. (2017) for Lawschool. The latter two are reported in Figure 2. For the training of CNF we kept the same hyperparameters used by Javaloy et al. (2023) for each of the three datasets: 1000 epochs, batch size of 256, and inner dimension of $[32, 32, 32]$.

Yet, we stress that our method does not train to generate counterfactual examples but only requires them as input, and may be employed in conjunction with any counterfactual generation methodology. These generation methods can be applied in a pre-processing step and the generated counterfactuals can then serve as input into our pipeline.

6.3 LTN Setup

As predicate for prediction in LTN, we train a multi-layer perceptron (MLP) with two layers of 100 and 50 neurons trained with the Adam optimizer with learning rate 0.1. We report averaged results over a 5-fold cross-validation. For LTN, we use Reichenbach implication and $p = 1$ as universal quantifier's exponent as introduced in Section 2.4. We report results for equally weighted axioms in the main paper and imbalanced weight settings in Appendix B. We ran all experiments on a computer with specification Ubuntu 22.04.1 LTS, 64 GB RAM and Ryzen Threadripper 1920X 12-Core Processor as CPU. Running times ranged from 1 minute to 1.5 hours for the largest dataset.

6.4 Baselines

In this section, we provide more information on each of the three baselines our pipeline is compared to: GAN-based method, DCEVAE and CNF. Also, we report hyperparameter settings and adaptions made for the comparison. All the following methodologies, differently from ours, generate counterfactual examples themselves. Our approach, however, is agnostic to the underlying counterfactual generation technique and may be easily integrated in existing pipelines that generate and extract counterfactual examples. This presents a challenge in terms of comparison, as these methods will tend to perform better on the set of counterfactuals that they themselves generated compared to other methodologies. Hence, we provide an evaluation of each baseline on a test set of the counterfactuals (approximated by CNF) we input into our pipeline (results in Section 7) as well as a study on how our method performs when we input the counterfactuals

generated by the GAN method (Appendix A). Furthermore, for DCEVAE and CNF, we train an MLP with the same hyperparameters as the underlying MLP in our method on the complete set of counterfactual and original data points. This is not to be confused with other proposed settings in literature (Javaloy et al. 2023), where predictors are sometimes trained in an *unaware* setting, which means that sensitive attributes are left out during training or only trained on non-descendent variables of the sensitive attribute.

GAN-based method. Grari et al. (2023) introduce a Generative Adversarial Model (GAN) approach for counterfactual inference and learning a counterfactually fair predictive model. For counterfactual inference, they propose a neural network encoder which generates a counterfactual from input X (original data point), Y and sensitive attribute S and a decoder which tries to reconstruct original Y and X from the generated data point and S. The adversarial network tries to infer S in this setting. For the counterfactual predictive model, they add an additional term for penalizing counterfactual unfairness to their loss function and extend this method to continuous features. We ran the available code[5] for 100 epochs for counterfactual inference and 1000 epochs for training a counterfactual fair predictor with learning rate 0.0001. For batch size we evaluated $[256, 512, 2048]$. Results are shown for 512 for Lawschool and COMPAS and 2048 for Adult. All other hyperparamters were set as given in the code.

DCEVAE. The Disentangled Causal Effect Variational AutoEncoder (DCEVAE) was proposed by Kim et al. (2021) as an extension to existing methodologies in fair variational optimisation. The main improvement put forward by the authors is the development of a ELBO-like objective for a causal graph in which variables that descend from sensitive attributes are kept separate from other covariates. The model then seeks to disentangle the VAE representations to separate the effects of the two sets of features. Among other applications, the authors test the counterfactual effect of applying their method to the Adult dataset.

In terms of integration into our experimental analysis, we started from the public code release by Kim et al. (2021)[6]. However, we noticed that the main PyTorch backprop code consistently gave a tensor version mismatch error. Thus, we modified the backprop loop by slightly changing the parameter update logic. We note that other authors that sought to reproduce the results from Kim et al. (2021) relied on the same bugfix.[7]

For hyperparameters we tested $[100, 250, 500]$ as training epochs and $[0.001, 0.0001]$ as learning rate as well as $[512, 1024]$ as batch size for all three datasets. We reported best results (100 epochs, 0.0001 learning rate, 1024 batch size) averaged over five runs.

CNF. We provide details on how we applied Causal Normalizing Flows (Javaloy et al. 2023) to approximate counterfactuals in Section 2.2. Provided these counterfactuals, we train an MLP with the same parameters as for our methodology on the combined dataset of counterfactuals and original data points. In terms of comparison, this baseline is the closest to our pipeline, as the same counterfactual generation method is applied. Yet, training is done differently as our method integrates counterfactual fairness constraints directly into the training pipeline and does not only take the generated counterfactual as input.

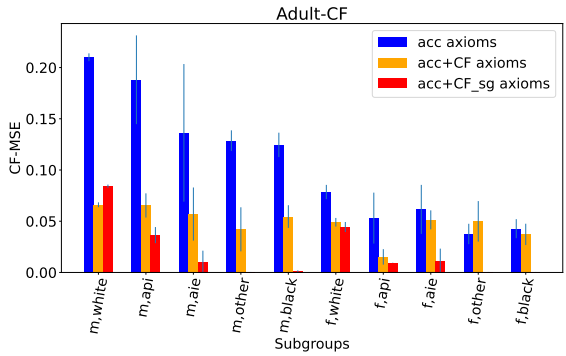


Figure 3. CF-MSE for the Adult dataset in three different axiom settings for each subgroup in (*gender, race*). Male corresponds to *m*, female to *f*, and *asian-pac-islander, american-indian-eskimo* are abbreviated with *api* and *aie*, respectively.

6.5 Evaluation Metrics

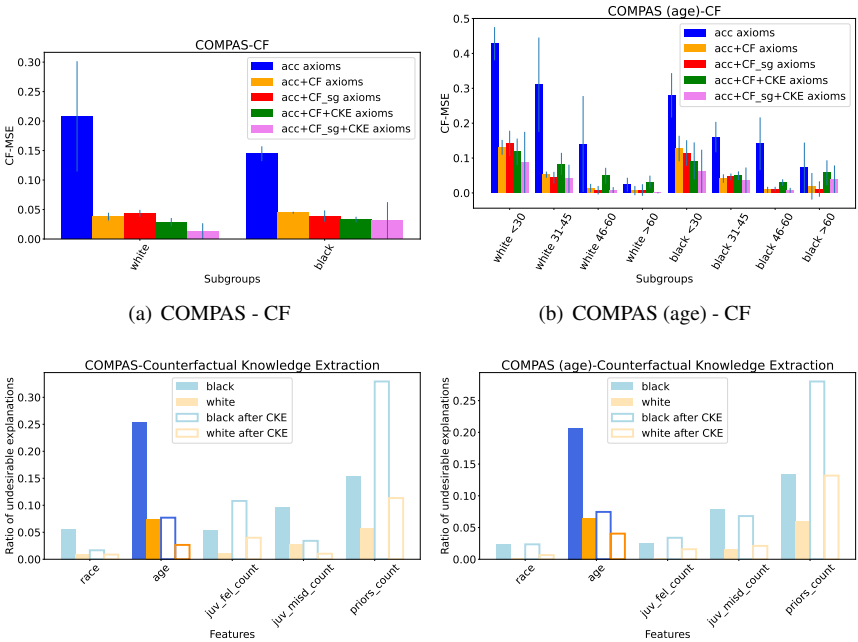
To assess the performance of our method and baselines, we employ metrics to measure both the predictive accuracy and the counterfactual fairness of the models across the diverse subgroups within each dataset.

The primary metric for evaluating counterfactual fairness is the **Counterfactual Mean Squared Error (CF-MSE)**. This measure quantifies the disparity in predictions between factual data points and their corresponding counterfactual instances. Specifically, CF-MSE is calculated as the average of the squared differences between the model's predictions for each factual data point, $D(x)$, and its counterfactual counterpart, $D(x')$, across the entire test set $\mathcal{T}$:

$$\text{CF-MSE} = \frac{1}{n} \sum_{x \in \mathcal{T}, x' \in cf(x,S)} |D(x) - D(x')|^2$$

Lower values of CF-MSE indicate a more counterfactually fair model, as this suggests minimal differences in predictions for individuals and their counterfactuals. To further evaluate the fairness of our method across different demographic groups, we introduce the **worst subgroup CF-MSE**. This metric highlights the maximum CF-MSE value observed across all evaluated subgroups (e.g., combinations of gender and race). By focusing on the worst-case scenario, we can assess the model's fairness in the most challenging contexts and ensure that our approach improves counterfactual fairness across all subgroups.

We also report the **accuracy** for the Adult and COMPAS datasets, and the **MSE** for the Lawschool dataset, to provide insight into the predictive performance of the models.



(a) COMPAS - CF

(b) COMPAS (age) - CF

(c) COMPAS - Undesirable Expl. Ratio

(d) COMPAS (age) - Undesirable Expl. Ratio

Figure 4. Top: Development of CF-MSE (lower is better) for our pipeline for 5 different axioms settings for both COMPAS datasets. Bottom: Ratio of undesirable explanations for each sensitive group before and after applying a CKE axiom for *(black, age)*.

7 Experimental Results

(Q1) How does our method improve counterfactual fairness, overall and at the subgroup level?

Setup. To show the effectiveness of our method, we evaluate three different axiom settings. As a baseline, we only apply the accuracy axioms (Equation A1-A2 or A3) to our pipeline (acc axioms). Next, we integrate the CF axiom (Equation A4) in addition to the accuracy axioms (acc+CF axioms). Lastly, we evaluate on the combination of subgroup axioms (Equation $A4_1$-$A4_n$) and accuracy axioms (acc+CF_sg axioms). All settings employing fairness axioms, pre-train an LTN for 1500 epochs on the accuracy axioms, then add the CF axioms.

Results. Results for **Adult** are displayed in Figure 3. There, one can see that applying the CF axioms strongly increases fairness for the majority of subgroups. The greatest improvement in CF-MSE can thereby be seen for the largest subgroup, namely *white males*, whereas for the female subgroups fairness only improves slightly and even gets worse for *females* in the "other" ethnic subgroup. However, integration of the subgroup axioms into the training objective mostly prevents this phenomenon. Overall,

Table 1. Comparison of our pipeline (three different axiom settings) with current baselines evaluated on CNF approximated counterfactuals in terms of accuracy, CF-MSE and worst subgroup CF-MSE (sg) as average of 5 runs. Row-wise best results are in bold.

		LTN (our pipeline)					
dataset	metric	acc	acc+CF	acc+ CF_sg	CNF	GAN	DCEVAE
Adult	accuracy ↑	0.782±0.006	0.758±0.01	0.812±0.001	0.825±0.001	0.777±0.005	**0.831±0.002**
	CF-MSE ↓	0.160±0.006	0.065±0.002	**0.055±0.001**	0.074±0.009	0.216±0.011	0.109±0.009
	CF-MSE (sg)↓	0.210±0.004	**0.066±0.002**	0.084±0.044	0.113±0.000	0.263±0.018	0.291±0.013
COMPAS	accuracy↑	0.671±0.010	0.675±0.003	0.651±0.013	0.661±0.003	**0.685±0.002**	0.665±0.009
	CF-MSE↓	0.156±0.037	0.047±0.004	**0.045±0.006**	0.072±0.010	0.107±0.012	0.188±0.051
	CF-MSE(sg)↓	0.208±0.094	0.045±0.002	**0.043±0.092**	0.086±0.010	0.110±0.012	0.194±0.061
COMPAS (age)	accuracy↑	0.658±0.013	0.654±0.016	0.658±0.012	0.667±0.002	**0.675±0.001**	0.651±0.008
	CF-MSE↓	0.254±0.032	0.079±0.016	**0.075±0.016**	0.094±0.003	0.171±0.010	0.244±0.044
	CF-MSE(sg)↓	0.428±0.048	**0.131±0.021**	0.142±0.057	0.181±0.006	0.204±0.013	0.342±0.074
Lawschool	MSE↓	0.767±0.013	0.782±0.002	0.796±0.013	0.771±0.008	0.906±0.000	**0.754±0.024**
	CF-MSE↓	0.096±0.028	0.003±0.000	**0.001±0.000**	0.012±0.002	0.227±0.013	0.210±0.042
	CF-MSE(sg)↓	0.358±0.021	0.011±0.002	**0.001±0.002**	0.014±0.003	0.272±0.025	0.251±0.046

CF improves for all subgroups upon the accuracy-only baseline; for all subgroups but white males, the CF-MSE is again improved by adding subgroup CF axioms. The same holds for both **COMPAS** datasets. In the top row plots of Figure 4, we show that CF axioms as well as subgroup CF axioms improve CF-MSE over all subgroups. We give complete numerical results for our LTN pipeline on all considered datasets in Table 1. Therein, we include average CF-MSE and worst-subgroup CF-MSE for all datasets and methods considered. The clear trend is that the average CF-MSE across groups improves when applying subgroup fairness axioms for all datasets. Also, accuracy is improved for Adult and both COMPAS datasets when applying subgroup CF axioms instead of the general CF axiom. For **Lawschool**, in Figure 5 we can see a huge improvement of CF for this dataset when adding CF axioms and an even stronger improvement when adding CF subgroup axioms that ensure counterfactual fairness with respect to *gender* to be distributed more evenly between the different races.

We conclude, in terms of **Q1**, that our methodology has a positive impact in terms of fairness, especially when subgroups are actively considered.

(Q2) How does our method compare to other approaches in terms of fairness and accuracy?

Setup. As baselines we compare our mehtod to DCEVAE (Kim et al. 2021), causal normalizing flows (CNF) (Javaloy et al. 2023) and a GAN-based method (Grari et al. 2023). For DCEVAE and CNF, we trained an MLP with the same hyper-parameters as our pipeline on the combined set of generated counterfactual and factual data points. Note that these methodologies, differently from ours, generate counterfactual examples themselves. This also means that they are not agnostic to the specific generated data points. To keep the comparison as fair as possible, we test all methodologies on the same counterfactual data, which we generate using a causal normalising flow model (Javaloy et al. 2023). Results on differently generated counterfactuals are given in Appendix A.

Results. We provide a complete comparison in Table 1. Regarding the comparison with CNF, results show that our pipeline, which adds counterfactual fairness constraints during training, significantly improves CF compared to only pre-processing for fairness as done by CNF. However, except for **COMPAS**, this results in a decrease in accuracy compared

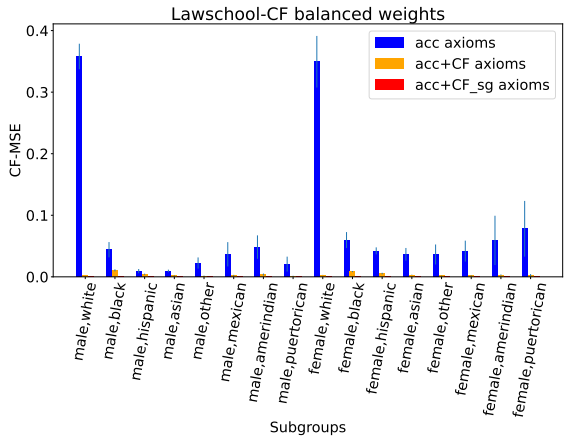


Figure 5. CF-MSE for Lawschool in three different axiom settings for each subgroup in (*gender, race*).

to CNF. For the GAN-based method, we can see improved accuracy for COMPAS but worse overall CF as well as worse subgroup CF in comparison to our method. Similarly, DCEVAE has strong results in terms of accuracy and MSE, but struggles in achieving counterfactually fair results. Here, our results differ significantly from the ones reported by the original authors (Kim et al. 2021). Our empirical, if anecdotal, experience with DCEVAE is that it struggles to converge to an accurate result, and even that comes at the expense of fairness. We elaborated on these reproducibility challenges in Section 6.4.

For **Q2**, we show that our technique has clear benefits, even if the comparison has some limitations as established above.

(Q3) Can counterfactual knowledge extraction be exploited to learn effective axioms?

Setup. Our pipeline, as described in Section 4.3, integrates counterfactual knowledge extraction (CKE henceforth). To summarise, CKE detects imbalances across sensitive groups in the frequency of undesirable counterfactual explanations, learning new training axioms to reduce them. To generate the counterfactual of data point x with respect to a generic immutable feature F (denoted by $cf(x, F)$) we use the method of causal normalizing flows described in Section 2.2. We define imbalance as a difference in frequencies of at least 0.1 for COMPAS, and 0.01 for Adult and Lawschool.

Results. We applied the extracted axioms both on top of models trained with the CF axiom (acc+CF+CKE) and with subgroup CF axioms (acc+CF_sg+CKE). Results for **COMPAS** are reported in Figure 4 as well as in Table 2. Here, the CKE deduced a strong imbalance for *age* for the *black* subgroups. After enforcing that age be irrelevant for decisions made in the $black$ group, the imbalance drops below the threshold. However, deincentivising *age* as counterfactual explanation results in the imbalance widening for

Table 2. Comparison of our proposed pipeline in five different axiom settings in terms of accuracy, CF-MSE and worst subgroup CF-MSE (sg) as average of 5 runs for both COMPAS datasets. CKE adds an axiom for *(black, age)*.

dataset	metric	acc	acc+CF	acc+CF_sg	acc+CF+CKE	acc+CF_sg+CKE
COMPAS	accuracy↑	0.671±0.010	**0.675±0.003**	0.651±0.013	0.645±0.008	0.633±0.022
	CF-MSE ↓	0.156±0.037	0.047±0.004	0.045±0.006	0.032±0.001	**0.025±0.006**
	CF-MSE(sg) ↓	0.208±0.094	0.045±0.002	0.043±0.092	0.033±0.005	**0.031±0.031**
COMPAS (age)	accuracy ↑	**0.658±0.013**	0.654±0.016	0.658±0.012	0.636±0.017	0.649±0.009
	CF-MSE ↓	0.254±0.032	0.079±0.016	0.075±0.016	0.077±0.027	**0.047±0.016**
	CF-MSE(sg)↓	0.428±0.048	0.131±0.021	0.142±0.057	0.118±0.038	**0.088±0.088**

other attributes – especially *priors_count* (Figures 4(c) and 4(d)). In the same figure, we observe that CF increases for both ethnicity subgroups even though axioms are only added for the *black* subgroup. A further increase of CF is achieved in the combination of subgroup CF axioms and the CKE axiom (Figures 4(a) and 4(b)). This can also be seen in Table 2, where CF-MSE is greatly reduced overall but also for subgroups when adding CKE axioms to disincentivise *age* for the *black* subgroup. Yet, we see a trade-off between improved CF-MSE and a loss in accuracy when applying additional CKE axioms. For **Adult**, we refer to Table 3 where we show how subsequently adding the detected CFK axioms influences the results. Therein, *race, marital-status, native-country* were interestingly detected as undesirable explanations for *males*. Here, for each detection a CKE axiom was added subsequently after 500 additional training epochs (ordered from highest imbalance to smallest imbalance), after which we each checked the axiom's impact on accuracy, CF-MSE and worst subgroup CF-MSE. Due to the axiom ordering by imbalance level, we have different sequences in which axioms are added for each run. Overall we found that accuracy stays stable throughout the CKE process. However, while CF is improved, subgroup CF gets worse with each additional CKE axiom after the first. It is left for further research how to establish scalability of the method beyond a single CKE axiom.

For **Lawschool**, as the counterfactual knowledge extraction works on binary predictions, we map the best 40% of all scores to a positive outcome. As a result, for Lawschool for one out of five runs *race* was detected as undesirable explanation for the *female* subgroup when CKE was evaluated after only training with the accuracy axioms. Yet, when CF axioms were added *race* was not detected as undesirable explanation anymore. Therefore, we conclude that CF axioms in this setting already eliminate undesirable explanations efficiently enough.

Our takeaway on the CKE technique, **Q3**, is that it is indeed able to learn beneficial axioms that reduce specific unfairness patterns for certain subgroups and feature combinations.

(Q4) What can we learn from post-hoc queries?

Setup. A key advantage of employing a neurosymbolic method, such as LTN, throughout our pipeline, lies in the opportunity for insights given by first order logic, post-training queries, an idea we discussed in Section 4.4. We showcase this capability by asking

Table 3. Impact for Adult of continuously adding CKE axioms on accuracy, CF-MSE and worst subgroup CF-MSE (sg). The CKE axioms are iteratively added in the order in which they appear in the table from left to right. Results are for one run, as the order of axioms varies across runs.

metric	LTN(acc)	LTN(acc+CF)	LTN (acc+CF+CKE)		
			(race,male)	*(mar.-status,male)*	*(nat.-country,male)*
accuracy ↑	0.778	0.755	0.757	0.755	0.755
CF-MSE ↓	0.183	0.063	0.052	0.058	0.050
CF-MSE (sg) ↓	0.224	0.073	0.070	0.100	0.139

a recourse-flavored query. That is, we take the perspective of a user that has received some undesired outcome (say, their loan request was denied) as a result of the model being employed. A natural fairness and recourse-related question for the user is then to ask whether there are *similar users in their subgroups which received a different outcome*, the same query we formalized earlier in Equation 2. The rationale here is to understand whether individuals in similar circumstances might have still succeeded under the model's resource-assignment rationale.

Results. In Table 4, we present illustrative how examples of this query's output in the **Adult** dataset. Therein, we show the query result for exemplary datapoints in each of four subgroups formed by the *gender, race* attribute columns. The results show that exemplary data points differ in the features *age, education-level, marital-status, relationship* and *occupation*. In particular, we observe for the white male example that the query returns a remarkably similar sample which differs only in age and marital status: The sample with a positive prediction (in **Adult**, a salary higher than $50\text{k}\$$) is 24 and married rather than 23 and unmarried. These discrepancies offer insights for individuals to examine the model's fairness and potentially challenge its decisions; on the model owner side, they are useful to glean more insights on the model's internal reasoning.

Overall, we conclude for **Q4** that a fair neurosymbolic method contributes to a wide range of additional knowledge extraction opportunities, enhancing understanding of the underlying data and learning process. This can not only strengthens the fairness guarantees of the model but also provides interpretable insights into its behaviour, fostering trust and transparency.

8 Conclusion & Future Work

To conclude, we have shown how to integrate the individual-based notion of counterfactual fairness into an LTN training pipeline. We proposed axioms for this integration and refined these axioms to subgroups, achieving higher counterfactual fairness for these subgroups by this. Furthermore, we integrated counterfactual knowledge extraction into our pipeline with subsequent axiom extraction to discourage undesirable counterfactual explanations. After training our model can be post-hoc queried for further information. Our pipeline improves counterfactual fairness and decreases the discrepancy between subgroups w.r.t. the unfair baseline, it has clear benefits over existing approaches and through its additional knowledge extraction

Table 4. Satisfaction value (sat) and exemplary data point to the query *is there a similar point in my subgroup which has a different outcome?* Here, τ is equal to 3 for males and 5 for females.

subgroup	sat	age	workclass	ed.	marital-status	occupation	relationship	h/w	nat.-country
white males	0.217	23	state-gov	12	never-married	adm-clerical	not-in-family	39	US
		24	state-gov	11	married-civ-spouse	adm-clerical	husband	39	US
black males	0.262	37	private	6	married-civ-spouse	handlers-cleaners	husband	39	US
		36	private	8	married-civ-spouse	machine-op-inspct	husband	39	US
white females	0.270	33	private	8	separated	adm-clerical	unmarried	39	US
		33	private	11	divorced	adm-clerical	not-in-family	39	US
asian-pac-isl. females	0.366	17	private	12	never-married	exec-managerial	other-relative	39	Philippines
		16	private	12	married-civ-spouse	exec-managerial	other-relative	39	China

opportunities enhances the understanding of the underlying data and learning process. This paper and the previous work we relate to suggest that the neurosymbolic approach to fairness is promising. It allows for the explicit and transparent codification of fairness axioms, but also potentially balance different axioms depending on the trade-offs/constraints for the application at hand.

Our work lays a foundation for further exploration at the intersection of neurosymbolic methods and (counterfactual) fairness. Building upon the knowledge extraction capabilities demonstrated in Q4, one direction for future work involves extending our pipeline to provide individuals with actionable recommendations. Specifically, we aim to identify concrete feature changes an individual can make to alter a potentially unfavourable model outcome. Further research could also focus on dynamically adjusting axiom weights based on performance and fairness considerations. As another addition to our setup, a future work should add experiments that focus on handling the issue of intersectionality, i.e. multiple interacting sensitive attributes, with LTN and fairness axioms. To our knowledge, our proposal is the first to explore the integration of counterfactual fairness principles with neurosymbolic architectures. Hence, the integration with other architectures is another research blind spot. As of now, the flavours of demographic parity, disparate impact, and counterfactual fairness have been formalized and implemented symbolically. Future research might target the formalization and efficient application of other notions, e.g. equalized odds (Hardt et al. 2016) or the Lipschitz condition as described by Dwork et al. (2012). Finally, we intend to evaluate our pipeline on more recent datasets, such as the ACS Income dataset (Ding et al. 2021), to provide additional tests of our method's generalizability. This direction of further studies could be extended by a practical case study.

Acknowledgements

XH and MC were supported by the "TOPML: Trading Off Non-Functional Properties of Machine Learning" project funded by Carl Zeiss Foundation, grant number P2021-02-014. Author VB was supported by a Royal Society University Research Fellowship.

Notes

1. The method takes the most frequent (max 10) values that are at least present in 1% of the training set. For continuous features, it takes the percentiles.
2. Limiting ourselves to those interventions on one feature only, which result in a change of the original prediction. Our interventions set the feature to the most frequent (max 10) values that are at least present in 1% of the training set. For continuous features, we take the percentiles.
3. Available at `https://github.com/psanch21/causal-flows`
4. Following Javaloy et al. (2023), we do not need to model the causal dependencies between the predictors and the target variable.
5. From `https://github.com/fairml-research/Counterfactual_Fairness`
6. available at `https://github.com/aailabkaist/DCEVAE`

7. For details, we refer to the `train.py` script on both the original repository, given above, and the following repository `https://github.com/osu-srml/CF_Representation_Learning/blob/master/DCEVAE/train.py`

References

Angwin J, Larson J, Mattu S and Kirchner L (2016) Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica* 23: 77–91.

Badreddine S, d'Avila Garcez A, Serafini L and Spranger M (2022) Logic tensor networks. *Artificial Intelligence* 303: 103649. DOI:10.1016/j.artint.2021.103649. URL `http://dx.doi.org/10.1016/j.artint.2021.103649`.

Becker B and Kohavi R (1996) Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Bhuyan BP, Ramdane-Cherif A, Tomar R and Singh TP (2024) Neuro-symbolic artificial intelligence: a survey. *Neural Computing and Applications* DOI:10.1007/s00521-024-09960-z. URL `http://dx.doi.org/10.1007/s00521-024-09960-z`.

Castelnovo A, Crupi R, Greco G and Regoli D (2021) The zoo of fairness metrics in machine learning. *CoRR* abs/2106.00467. URL `https://arxiv.org/abs/2106.00467`.

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

Deck L, Schoeffer J, De-Arteaga M and Kühl N (2024) A critical survey on fairness benefits of explainable ai. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505, p. 1579–1595. DOI:10.1145/3630106.3658990. URL `https://doi.org/10.1145/3630106.3658990`.

Ding F, Hardt M, Miller J and Schmidt L (2021) Retiring adult: New datasets for fair machine learning. In: Ranzato M, Beygelzimer A, Dauphin YN, Liang P and Vaughan JW (eds.) *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual.* pp. 6478–6490. URL `https://proceedings.neurips.cc/paper/2021/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html`.

Dwork C, Hardt M, Pitassi T, Reingold O and Zemel R (2012) Fairness through awareness. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12. New York, NY, USA: Association for Computing Machinery. ISBN 9781450311151, p. 214–226. DOI:10.1145/2090236.2090255. URL `https://doi.org/10.1145/2090236.2090255`.

Gibaut W, Pereira L, Grassiotto F, Osorio A, Gadioli E, Munoz A, Gomes S and Santos Cd (2023) Neurosymbolic ai and its taxonomy: a survey DOI:10.48550/ARXIV.2305.08876. URL `https://arxiv.org/abs/2305.08876`.

Goethals S, Martens D and Calders T (2024) Precof: counterfactual explanations for fairness. *Machine Learning* 113: 3111–3142. DOI:10.1007/s10994-023-06319-8.

Grari V, Lamprier S and Detyniecki M (2023) Adversarial learning for counterfactual fairness. *Machine Learning* 112(3): 741–763.

Greco G, Alberici F, Palmonari M and Cosentini A (2023) Declarative encoding of fairness in logic tensor networks. *ECAI 2023* DOI:10.3233/faia230360. URL http://dx.doi.org/10.3233/FAIA230360.

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

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

Hitzler P and Sarker MK (2022) *Neuro-symbolic Artificial Intelligence: The State of the Art*. Frontiers in artificial intelligence and applications. IOS Press. ISBN 9781643682440. URL https://books.google.co.uk/books?id=jnL0zgEACAAJ.

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

Javaloy A, Sanchez-Martin P and Valera I (2023) Causal normalizing flows: from theory to practice. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M and Levine S (eds.) *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., pp. 58833–58864. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/b8402301e7f06bdc97a31bfaa653dc32-Paper-Conference.pdf.

Karimi AH, Barthe G, Schölkopf B and Valera I (2022) A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Comput. Surv.* 55(5). DOI:10.1145/3527848. URL https://doi.org/10.1145/3527848.

Kim H, Shin S, Jang J, Song K, Joo W, Kang W and Moon IC (2021) Counterfactual fairness with disentangled causal effect variational autoencoder. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35. pp. 8128–8136.

Kocaoglu M, Snyder C, Dimakis AG and Vishwanath S (2018) CausalGAN: Learning causal implicit generative models with adversarial training. In: *International Conference on Learning Representations*. URL https://openreview.net/forum?id=BJE-4xW0W.

Kusner MJ, Loftus J, Russell C and Silva R (2017) Counterfactual fairness. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S and Garnett R (eds.) *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Lin Y, Zhao C, Shao M, Meng B, Zhao X and Chen H (2024) Towards counterfactual fairness-aware domain generalization in changing environments. In: Larson K (ed.) *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*. International Joint Conferences on Artificial Intelligence Organization, pp. 4560–4568. DOI:10.24963/ijcai.2024/504. URL https://doi.org/10.24963/ijcai.2024/504. Main Track.

Louizos C, Shalit U, Mooij JM, Sontag D, Zemel R and Welling M (2017) Causal effect inference with deep latent-variable models. *Advances in neural information processing systems* 30.

Lundberg SM and Lee SI (2017) A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S and Garnett R (eds.) *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Pearl J and Mackenzie D (2018) *The Book of Why*. New York: Basic Books. ISBN 978-0-465-09760-9.

Russell C, Kusner MJ, Loftus J and Silva R (2017) When worlds collide: Integrating different counterfactual assumptions in fairness. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S and Garnett R (eds.) *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/1271a7029c9df08643b631b02cf9e116-Paper.pdf.

Sarker MK, Zhou L, Eberhart A and Hitzler P (2021) Neuro-symbolic artificial intelligence: Current trends.

Simson J, Fabris A and Kern C (2024) Lazy data practices harm fairness research. In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24. ACM. DOI:10.1145/3630106.3658931. URL http://dx.doi.org/10.1145/3630106.3658931.

Verma S and Rubin J (2018) Fairness definitions explained. In: *Proceedings of the International Workshop on Software Fairness*, FairWare '18. New York, NY, USA: Association for Computing Machinery. ISBN 9781450357463, p. 1–7. DOI:10.1145/3194770.3194776. URL https://doi.org/10.1145/3194770.3194776.

Wagner B and d'Avila Garcez AS (2021) Neural-symbolic integration for fairness in ai. In: *AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021)*, volume 2846. URL http://ceur-ws.org/Vol-2846/paper5.pdf. © 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

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

Wightman LF (1998) Lsac national longitudinal bar passage study. lsac research report series. URL https://api.semanticscholar.org/CorpusID:151073942.

Xu D, Wu Y, Yuan S, Zhang L and Wu X (2019) Achieving causal fairness through generative adversarial networks. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

Yang M, Liu F, Chen Z, Shen X, Hao J and Wang J (2021) Causalvae: Disentangled representation learning via neural structural causal models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 9593–9602.

Zhang L, Wu Y and Wu X (2016) Achieving non-discrimination in data release. URL https://arxiv.org/abs/1611.07438.

Zuo Z, Khalili M and Zhang X (2023) Counterfactually fair representation. *Advances in Neural Information Processing Systems* 36: 12124–12140.

A Comparing to Other Counterfactuals

As stressed before, unlike ours, all methodologies we compare to generate counterfactual examples themselves. Our method relies only on a set of counterfactuals given as input, so that it is agnostic to the underlying counterfactual generation technique. However, during comparison of the different methods we faced the challenge that just a comparison of methods without taking the generated counterfactuals into account is not appropriate for our method. We therefore firstly compared each baseline on a common test set of the counterfactuals (approximated by CNF) we input into our pipeline. Secondly, we took the counterfactuals generated by the GAN-based method as input into our method and compared it to the GAN pipeline. For this comparison, we had to modify the GAN-based method, as in the original version CF-MSE and accuracy is calculated on different data encodings which was not possible as input into our pipeline. In Table 5 the results show better values for CF-MSE when training with our method. For COMPAS and COMPAS(age) this results in a decreased accuracy, compared to the GAN-based method. However, for the Adult and Lawschool dataset accuracy and MSE is improved upon the GAN method. Altogether, these results show that our method is applicable to counterfactuals generated with different methods than with CNF. Also, for these counterfactuals our method shows improved results, specifically for CF-MSE, when compared to the original generation method.

B Influence of Axiom Weights

Our pipeline supports different weights for each group of axioms (accuracy, CF, CKE). This has direct influence on CF and accuracy as can be seen in Table 6. As a trend, accuracy improves, if higher weights are chosen for the accuracy axioms while CF decreases. However, this is not the case for all datasets, and we suggest here to try out different weight settings when applying our pipeline.

Table 5. Comparison of our proposed pipeline (with two different axiom settings) with the GAN baseline evaluated on the counterfactuals the GAN method produces in terms of accuracy, CF-MSE and worst subgroup CF-MSE (sg) as average of 5 runs. Best results are in bold.

dataset	metric	LTN(acc)	LTN(acc+CF)	GAN
Adult	accuracy ↑	0.771 ± 0.007	$\mathbf{0.775 \pm 0.005}$	0.758 ± 0.006
	CF-MSE ↓	0.231 ± 0.007	$\mathbf{0.200 \pm 0.007}$	0.208 ± 0.025
COMPAS	accuracy ↑	0.672 ± 0.015	0.658 ± 0.016	$\mathbf{0.680 \pm 0.001}$
	CF-MSE ↓	0.259 ± 0.016	$\mathbf{0.115 \pm 0.011}$	0.177 ± 0.015
COMPAS(age)	accuracy ↑	0.653 ± 0.013	0.654 ± 0.007	$\mathbf{0.668 \pm 0.006}$
	CF-MSE ↓	0.321 ± 0.011	$\mathbf{0.210 \pm 0.018}$	0.249 ± 0.042
Lawschool	MSE ↓	0.235 ± 0.002	$\mathbf{0.234 \pm 0.002}$	0.906 ± 0.001
	CF-MSE ↓	0.019 ± 0.015	$\mathbf{0.019 \pm 0.015}$	0.256 ± 0.028

Table 6. Comparison of our proposed pipeline with two different weight combinations for acc+CF axioms as well as for acc+CF_sg.

dataset	metric	LTN (acc+CF)		LTN (acc+CF_sg)	
		(1,1)	(2,1)	(1,1)	(2,1)
Adult	accuracy ↑	0.758±0.01	0.772±0.006	**0.812±0.001**	0.769±0.004
	CF-MSE ↓	0.065±0.002	0.078±0.002	**0.055±0.001**	0.066±0.001
	CF-MSE (sg) ↓	**0.066±0.002**	0.094±0.054	0.084±0.044	0.084±0.004
COMPAS	accuracy ↑	0.675±0.003	0.674±0.013	0.651±0.013	**0.683±0.009**
	CF-MSE ↓	0.047±0.004	0.054±0.007	**0.045±0.006**	0.052±0.002
	CF-MSE (sg) ↓	0.045±0.002	0.057±0.012	**0.043±0.092**	0.049±0.078
COMPAS(age)	accuracy ↑	0.654±0.016	0.653±0.011	**0.658±0.012**	0.655±0.018
	CF-MSE↓	0.079±0.016	0.114±0.022	0.075±0.016	**0.070±0.018**
	CF-MSE (sg)↓	**0.131±0.021**	0.269±0.091	0.142±0.057	0.143±0.078
Lawschool	MSE ↓	0.782±0.002	0.773±0.018	**0.796±0.013**	0.786±0.018
	CF-MSE ↓	0.003±0.000	0.005±0.000	**0.001±0.000**	0.002±0.000
	CF-MSE (sg) ↓	0.011±0.002	0.011±0.001	**0.001±0.002**	0.002±0.000