A Proof System with Causal Labels (Part I): checking Individual Fairness and Intersectionality

Leonardo Ceragioli[1,*,†], Giuseppe Primiero[1,*,†]

[1]University of Milan, Via Festa del Perdono 7, Milano, 20122, Italy

Abstract

In this article we propose an extension to the typed natural deduction calculus **TNDPQ** to model verification of individual fairness and intersectionality in probabilistic classifiers. Their interpretation is obtained by formulating specific conditions for the application of the structural rule of Weakening. Such restrictions are given by causal labels used to check for conditional independence between protected and target variables.

Keywords

Fairness, Probabilistic Logic, Causal Graphs, Structural Rules

1. Introduction

The calculus **TPTND** (*Trustworthy Probabilistic Typed Natural Deduction* [1, 2]) is designed to evaluate *post-hoc* the trustworthiness of the behavior of opaque systems. The system is implemented for verification of dataframes in the tool BRIO [3, 4]. In [5], we introduced **TNDPQ** (*Typed Natural Deduction for Probabilistic Queries*), a variation of the previous system in which a probabilistic output is associated to a target variable when a Data Point – consisting of a list of values attributions for a set of variables – is provided. In this paper, we extend this system with tools to verify individual fairness of classifiers.

We start with a formal definition of classifiers. Let $\mathscr{A}$ be a set of protected variables $a_1, \ldots, a_n$, $\mathscr{X}$ be a (disjoint) set of non-protected variables $x_1, \ldots, x_n$ and t be a target variable. Moreover, let $\mathscr{V}_{a_i}$ be a set of values $\alpha^{i_1}, \alpha^{i_2}, \ldots, \alpha^{i_j}$ that a_i can receive, $\mathscr{V}_A$ the set of all $\mathscr{V}_{a_i}$, $\mathscr{V}_{x_i}$ be a set of values $\beta^{i_1}, \beta^{i_2}, \ldots, \beta^{i_j}$ that x_i can receive, $\mathscr{V}_X$ the set of all $\mathscr{V}_{x_i}$, and $\mathscr{V}_t$ the set $\delta^1, \delta^2, \ldots, \delta^j$ of values that t can receive. Let us use $v_1, \ldots, v_n$ to denote elements of $\mathscr{A} \cup \mathscr{X}$ (that is, variables regardless of their protected or unprotected status), and $\gamma^{i_1}, \ldots, \gamma^{i_j}$ to denote the values that v_i can receive. We use $a_i : \alpha^{i_j}$ (respectively $x_i : \beta^{i_j}$) to express the *judgment* that variable a_i receives value α^{i_j} (respectively, variable x_i receives value β^{i_j}), and $t : \delta^1_{p_1}, \ldots, \delta^j_{p_j}$ to express the *probabilistic judgment* that $\delta^1, \ldots, \delta^j$ are all the possible values that variable t can receive and that, for $1 \leq k \leq j$, it receives value δ^k with probability p_k.[1] We use $\mathscr{J}^{\mathscr{A}}$ for the set of all the judgments about protected variables, $\mathscr{J}^{\mathscr{X}}$ for the set of all the judgments about non-protected variables, and $\mathscr{J}^{\mathscr{P}}$ for the set of all probabilistic judgments. Moreover, we use $\sigma^{\mathscr{A}}$ to express a set of judgments about protected variables such that each element of $\mathscr{A}$ receives at most one value, $\sigma^{\mathscr{X}}$ to express a set of judgments about non-protected variables such that each element of $\mathscr{X}$ receives at most one value, $\Sigma^{\mathscr{A}}$ to refer to the set of all $\sigma^{\mathscr{A}}$, and $\Sigma^{\mathscr{X}}$ to refer to the set of all $\sigma^{\mathscr{X}}$. σ is used to express the union of a $\sigma^{\mathscr{A}}$ and a $\sigma^{\mathscr{X}}$, and Σ is used to refer to the set of all σ. More formally:

$$\Sigma^{\mathscr{A}} =_{def} \{\sigma^{\mathscr{A}} \subseteq \mathscr{J}^{\mathscr{A}} \mid \forall i(a_i : \alpha^{i_l} \in \sigma^{\mathscr{A}} \wedge a_i : \alpha^{i_m} \in \sigma^{\mathscr{A}} \rightarrow l = m)\}$$

7th International Workshop on Artificial Intelligence and fOrmal VERification, Logic, Automata, and sYnthesis , OVERLAY 2025, Oct 25 – 26, Bologna, Italy

*Corresponding author.

†These authors contributed equally.

leonardo.ceragioli@unimi.it (L. Ceragioli); giuseppe.primiero@unimi.it (G. Primiero)

0000-0001-5250-9720 (L. Ceragioli); 0000-0003-3264-7100 (G. Primiero)

[1]We assume that the values for t (also for the elements of $\mathscr{A}$ and $\mathscr{X}$ but this is irrelevant here) are all mutually exclusive, and so $\sum_{k=1}^{j} p_k = 1$. Note that we make no assumption regarding whether $t \in \mathscr{A} \cup \mathscr{X}$ and so on whether $\mathscr{V}_t = \mathscr{V}_{a_i}$ or $\mathscr{V}_t = \mathscr{V}_{x_i}$ for some i.

Table 1
The table shows the 680 Data Points in the Training Set that satisfy σ. Of them: 100 satisfy $a_1 : \alpha^{1_1}$ and $a_2 : \alpha^{2_1}$, 90 of which satisfy also $t : \delta$; 240 satisfy $a_1 : \alpha^{1_1}$ and $a_2 : \alpha^{2_2}$, 180 of which satisfy also $t : \delta$; 240 satisfy $a_1 : \alpha^{1_2}$ and $a_2 : \alpha^{2_1}$, 180 of which satisfy also $t : \delta$; 100 satisfy $a_1 : \alpha^{1_2}$ and $a_2 : \alpha^{2_2}$, 90 of which satisfy also $t : \delta$. By summing up the points in each column and in each row we obtain that 79% of the points satisfy $t : \delta$, so the lack of fairness disappears when at most one of a_1 or a_2 is considered.

		attribute a_1		
		α^{1_1}	α^{1_2}	
attribute a_2	α^{2_1}	$\frac{90}{100} \approx 0.90$	$\frac{180}{240} \approx 0.75$	$\frac{270}{340} \approx 0.79$
	α^{2_2}	$\frac{180}{240} \approx 0.75$	$\frac{90}{100} \approx 0.90$	$\frac{270}{340} \approx 0.79$
		$\frac{270}{340} \approx 0.79$	$\frac{270}{340} \approx 0.79$	

$$\Sigma^{\mathscr{X}} =_{def} \{\sigma^{\mathscr{X}} \subseteq \mathscr{J}^{\mathscr{X}} \mid \forall i(x_i : \beta^{i_l} \in \sigma^{\mathscr{X}} \wedge x_i : \beta^{i_m} \in \sigma^{\mathscr{X}} \rightarrow l = m)\}$$
$$\Sigma =_{def} \{\sigma^{\mathscr{A}} \cup \sigma^{\mathscr{X}} \mid \sigma^{\mathscr{A}} \in \Sigma^{\mathscr{A}} \ \wedge \ \sigma^{\mathscr{X}} \in \Sigma^{\mathscr{X}}\}$$

A classifier $\widehat{f} \in \widehat{\mathscr{F}}$ is a function from Σ to $\mathscr{J}^{\mathscr{P}}$, where each $\sigma \in \Sigma$ describes a Data Point, that is what we know about a subject, and the probabilistic judgment $t : \delta^1_{p_1}, \ldots, \delta^j_{p_j}$ in $\mathscr{J}^{\mathscr{P}}$ represents the output of the classifier regarding the probability distribution of the possible values for the target variable t.

TNDPQ is a proof system working with sequents describing the result of queries for classifiers. More precisely, each classifier $\widehat{f}$ is characterized by a set of ground sequents of the form:[2]

$$\sigma \mid\sim \ t : \delta_p \tag{1}$$

For readability reasons, the sequents focus on only one possible value for the target variable at a time. In [5] we show how to extend **TNDPQ** with sequents working with logically complex judgments – possibly non-atomic variables receiving possibly non-atomic values. As an example, the following sequent expresses the probability that a non-white 27 years old woman who is married or divorced receives a loan:

$$Age : 27,\ Gen. : f,\ MS : married + divorced,\ Etn. : white^{\perp} \ \Big|\sim \ Loan : yes_{0.60}$$

TNDPQ was initially designed to investigate the preservation of trustworthiness under the composition of logically simpler queries. In this paper, we focus only on the atomic fragment and provide a *criterion* to verify individual fairness for a probabilistic classifier via structural properties. Moreover, we address the issue of intersectionality, showing a solution through an extension of the original language with causal relations.

2. Individual Fairness

Individual fairness has different non-equivalent definitions in the literature. A first simplified characterization of *individual fairness* is as follows:

Definition 2.1 (Individual Fairness (**IF**))**.** A classifier is individually fair regarding a set of protected attributes if it gives the same outputs to Data Points differing only for the values of those attributes. Formally, $\widehat{f}$ is **IF** regarding the set of protected attributes $\{a_1, \ldots, a_n\}$ iff for every $\sigma^{\mathscr{X}} \in \Sigma^{\mathscr{X}}$ and pair of n-tuples of values $\alpha^{1_l}, \ldots, \alpha^{n_l}$ and $\alpha^{1_m}, \ldots, \alpha^{n_m}$, $\widehat{f}(\sigma^{\mathscr{X}}, a_i : \alpha^{1_l}, \ldots, a_i : \alpha^{n_l}) = \widehat{f}(\sigma^{\mathscr{X}}, a_i : \alpha^{1_m}, \ldots, a_i : \alpha^{n_m})$.

[2]Technically, we should add a subscript in the equation specifying the classifier we are focusing on. However, this will not be needed here, since we will not compare outputs of different classifiers.

Given definition 2.1, in **TNDPQ** a classifier is **IF** regarding the set of protected attributes $\{a_1, \ldots, a_n\}$ iff the ground sequents describing the classifier are such that $\sigma^{\mathscr{X}}, a_i : \alpha^{1_l}, \ldots, a_i : \alpha^{n_l} \mid\sim\ t : \delta_p$ iff $\sigma^{\mathscr{X}}, a_i : \alpha^{1_l}, \ldots, a_i : \alpha^{n_l} \mid\sim\ t : \delta_p$, for every $\delta \in \mathscr{V}_t$, $\alpha^{i_l}, \alpha^{i_m} \in \mathscr{V}_{a_i}$, and $\sigma^{\mathscr{X}} \in \Sigma^{\mathscr{X}}$.

We consider **IF** as the best way of approximating *fairness through unawareness* when we have to evaluate opaque systems:

Definition 2.2 (Fairness through Unawareness (**FtU**))**.** A classifier is fair through unawareness regarding a protected attribute as long as this attribute is not explicitly used in the decision-making process.

Indeed, while **FtU** is clearly an *intensional* notion, which can be properly evaluated only by looking at the implemented program, **IF** can be evaluated just by looking at the inputs and outputs of the machine. For this reason, we cannot directly evaluate **FtU** for opaque systems, and **IF** emerges as a good substitute.

However, there is a problem. Although intersectionality holds for **FtU**, it fails for **IF**.[3]

Observation 2.1 (Intersectionality fails for **IF**)**.** *A classifier that is **IF** with respect to protected attribute a_1 and protected attribute a_2 (separately) may not be **IF** regarding the set $\{a_1, a_2\}$.*

Proof. To prove failure of intersectionality, we just have to show that for some classifier $\widehat{f}$, for every $\sigma^{\mathscr{X}} \in \Sigma^{\mathscr{X}}$, for every pair of values α^{1_l} and α^{1_m} in $\mathscr{V}_{a_1}$, and for every pair of values α^{2_l} and α^{2_m} in $\mathscr{V}_{a_2}$

$$\widehat{f}(\sigma^{\mathscr{X}}, a_1 : \alpha^{1_l}) = \widehat{f}(\sigma^{\mathscr{X}}, a_1 : \alpha^{1_m})$$
$$\widehat{f}(\sigma^{\mathscr{X}}, a_2 : \alpha^{2_l}) = \widehat{f}(\sigma^{\mathscr{X}}, a_2 : \alpha^{2_m})$$

but for some pair of sets of values $\alpha^{1_l}, \alpha^{2_l}$ and $\alpha^{1_m}, \alpha^{2_m}$

$$\widehat{f}(\sigma^{\mathscr{X}}, a_1 : \alpha^{1_l}, a_2 : \alpha^{2_l}) \neq \widehat{f}(\sigma^{\mathscr{X}}, a_1 : \alpha^{1_m}, a_2 : \alpha^{2_m})$$

We will show that such a classifier is not only theoretically possible, but even quite common when ML systems are trained using Data Sets of a specific kind.

For simplicity, assume that a_1 and a_2 have only two possible outputs each (respectively α^{1_1} and α^{1_2}, and α^{2_1} and α^{2_2}). Let us consider an ML system implementing a learning algorithm with no restriction on protected attributes. The system is not **FtU** regarding these attributes, but can be **IF** if it is trained using a fair Data Set: that is, a Data Set in which all the Data Points sharing the same value of the non-protected attributes but possibly differing for those of a_1 or a_2 share the same value of the target variable. In our case, let us assume that the Data Set is fair in this sense and focus on the specific Data Points in table 1: these are all the Data Points that satisfy a specific $\sigma^{\mathscr{X}} \in \Sigma^{\mathscr{X}}$, with the ratio expressing how many of them give value δ to target variable t. We can observe, by looking at the table, that the Data Points are fair when a_1 and a_2 are considered separately, and biased when a_1 and a_2 are considered together. Hence, the ML system will not learn to be **IF** regarding $\{a_1, a_2\}$. As an example:

$$\sigma^{\mathscr{X}}, a_1 : \alpha^{1_1}, a_2 : \alpha^{2_1} \mid\sim\ t : \delta_{0.90}$$
$$\sigma^{\mathscr{X}}, a_1 : \alpha^{1_1}, a_2 : \alpha^{2_2} \mid\sim\ t : \delta_{0.75}$$

Note that a different probability associated with just one value of t is sufficient to disprove **IF** for the set of variables $\{a_1, a_2\}$ and so the proof is complete. □

Lack of intersectionality is even more serious than it could seem. Indeed, if intersectionality fails, fairness can be gerrymandered by cherry picking protected attributes, so this property contributes to make **IF** relevant even if the focus is only on single protected attributes [6, 7]. Moreover, intersectionality also fails for more elaborated notions of individual fairness which implement a metric for similarity of Data Points and require similar predictions for similar points [8, 9]. In fact, while a shared assumption in the existing literature about **IF** is that all features of the Data Points are mutually independent, in the next section we argue that the causal relations among such features must be taken into account in order to check intersectionality for opaque systems.

[3] Notice that, even though we focus only on the case of two protected variables, the result generalizes for any set of protected variables.

3. IF and intersectionality as Weakening

Since **IF** and intersectionality require that the probability of a sequent does not change when different values are attributed to protected variables, both these properties can be seen as equivalent to a restricted rule of *Weakening*:[4]

$$\frac{\sigma \mid\sim \ t : \delta_p}{\sigma, a : \alpha \mid\sim \ t : \delta_p} \ Weakening^* \tag{2}$$

Hence, since **TNDPQ** is non-monotonic, we need to provide the *condition* under which this rule is valid, in order to deal with **IF** and intersectionality. As an example, the following inference establishes **IF** regarding the protected attribute *gender* (variable $Gen.$) and an instance of intersectionality in case also *marital status* (MS) is considered protected:

$$\frac{Age : 27,\ MS : married + divorced,\ Etn. : white^{\perp} \mid\sim \ Loan : yes_{0.60}}{Age : 27,\ Gen. : f,\ MS : married + divorced,\ Etn. : white^{\perp} \mid\sim \ Loan : yes_{0.60}} \ Weakening^*$$

Since **TNDPQ** is a probabilistic system, *Weakening* is valid when the active variable in the rule (a) and the target variable (t) are mutually independent, conditional on σ.

Definition 3.1 (Conditional Independence)**.** t and a are independent, conditional on σ, iff

$$P(t : \delta \mid a : \alpha, \sigma) = P(t : \delta \mid \sigma)$$

Therefore, what we need is a *criterion* of conditional independence.

In a logic that ignores the relations between the elements of σ, conditional independence can be decided only using brute-force methods, i.e. by checking statistical correlations for all values of any variable. Moreover, we have no way of distinguishing good correlations from spurious ones which may emerge due to biases in the Training Set. In contrast, a system extended with the basic vocabulary of causal graphs [10] can describe directly causal relations between the features of the classifier. Hence, we translate in our calculus some well-studied conditions for independence, obtaining admissibility *criteria* for Weakening, in turn establishing **IF** and intersectionality. For the purposes of this work, we define causal graphs as follows:

Definition 3.2 (Causal Graph)**.** A causal graph is an acyclic directed graph with nodes representing events (variables receiving values) and edges representing immediate causal relations.

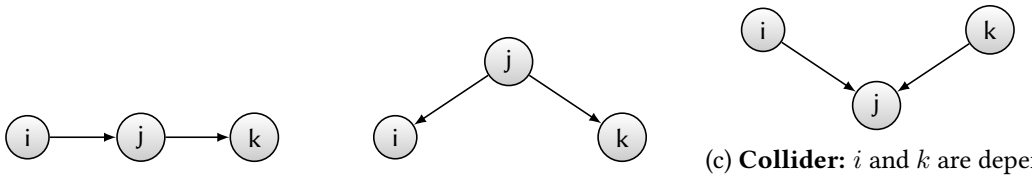
By closing edges under transitivity, we obtain the notion of mediate cause. For purely formal reasons, we close the notion of cause under reflexivity as well. The usual extension of deterministic causal graphs with functions to compute the value of a node on the basis of those of all the immediate parent nodes is here expressed by sequents like in equation 1. Moreover, the common distinction between exogenous and endogenous nodes, relevant in discussing interventions and counterfactual fairness, is left for further research.

Two nodes which are one the immediate cause of the other are mutually dependent. While for two nodes which are not directly connected, dependence is defined in three steps [10]:

- The *criteria* in figure 1 deal with the easiest cases possible, when there is only one intermediate node between the two;
- We define a path as blocked by a set of nodes iff, when all and only these nodes occur in the condition, at least one chain, fork or collider in the path has independent nodes;
- Two nodes are independent iff all the paths between them are blocked.

To express the *criteria* for conditional independence and thus for the applicability of the rule of Weakening, we need to internalize the causal notions in our calculus **TNDPQ**. For this purpose, we use the methodology of labeled calculi [11, 12]. First, we extend the language with the following relational predicates for variables:

[4]Notice that intersectionality is addressed by considering σ and not only $\sigma^{\mathscr{X}}$.



(a) **Chain:** i and k are dependent, but independent conditional on j.

(b) **Fork:** i and k are dependent, but independent conditional on j.

(c) **Collider:** i and k are dependent conditional on j or any of its descendant, independent otherwise.

Figure 1: Elementary compositions of nodes in causal graphs and *criteria* of conditional independence.

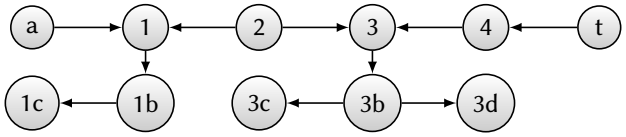


Figure 2: Paths between nodes a and t.

Immediate Causal Relations $v_i \triangleright v_j =_{def} v_i$ is an immediate cause of v_j.

Mediate Causal Relations $v_i \blacktriangleright^{M} v_j =_{def} v_i$ is a mediate cause of v_j, with intermediate nodes M.

Path with Intermediate Nodes $v_i \Diamond_N^M v_j =_{def}$ a path exists between v_i and v_j passing through non-colliders M and colliders or sets of their descendants N.

then, we reformulate **TNDPQ** sequents by extending their left-hand side with causal relations:

$$\triangleright_{\widehat{f}}, \sigma \left|\sim\ t : \delta_p \right. \tag{3}$$

Let us use Var_σ to indicate the set of variables that occur in σ. We use $\triangleright_{\widehat{f}}$ to indicate all the immediate causal relations among features in the classifier. $\Diamond_{\widehat{f}}$ denotes all the existing paths in the resulting graph and is derivable as the closure of $\triangleright_{\widehat{f}}$ under the rules in table 2.

To see how these rules work, consider the figure 2. The rules for mediate causal relations are used to identify the descendants of colliders. In this case, applying **reflexive cause** and **transitive cause** we obtain: $1 \blacktriangleright^{\{1,1b,1c\}} 1c$, $3 \blacktriangleright^{\{3,3b,3c\}} 3c$, and $3 \blacktriangleright^{\{3,3b,3d\}} 3d$. The rules **chain**, **fork**, and **collider** are used to represent triplets of nodes between a and t, with the rule for collider representing also the descendants. Chains and forks are stored in the superscript: $1 \Diamond^{\{2\}} 3$, $3 \Diamond^{\{4\}} t$. The colliders are stored in the subscript, together with the set of their descendants: $a \Diamond_{\{1,\{1,1b,1c\}\}} 2$, $2 \Diamond_{\{3,\{3,3b,3c\}\}} 4$, and $2 \Diamond_{\{3,\{3,3b,3d\}\}} 4$.[5] **Transitivity** is used to combine these triplets to construct the paths between a and t. In this case, we have two paths: $a \Diamond^{\{2,4\}}_{\{1,\{1,1b,1c\},3,\{3,3b,3c\}\}} t$, and $a \Diamond^{\{2,4\}}_{\{1,\{1,1b,1c\},3,\{3,3b,3d\}\}} t$.[6]

Note that in this calculus variables play the same role as labels in labeled calculi, with $\triangleright_{\widehat{f}}$ making explicit accessibility relations. As an example, the sequent expressing that the probability that a 27 years old person with a gross annual income of 40.000 receives a loan is 60%, is formulated as follows:

$$Age \triangleright MS, Age \triangleright GAI, Age \triangleright Loan, GAI \triangleright Loan, Age : 27,\ GAI : 40K \mid\sim\ Loan : yes_{0.60}$$

With these technical tools in place, we formulate an applicability *criterion* for the *Weakening* rule in equation 4:

[5]Notice that the collider occurs both by itself and in the set of its descendants: its occurrence as collider is needed for the rule of transitivity, and its occurrence in the set of its descendants is needed for the condition of applicability of the rule (4).

[6]We focus only on the maximal sets of descendants, although technically also the paths containing only some of the descendants are constructible. Notice that this does not cause problems with the conditions of rule (4).

Table 2
Rules to derive paths $\diamondsuit_{\widehat{f}}$ from immediate causal relations $\triangleright_{\widehat{f}}$. The rule of Transitivity has the condition that $i \in O \cup P$ and $j \in M \cup N$.

Reflexive cause $\vdash v_i \blacktriangleright^{\{i\}} v_i$

Transitive cause $v_i \blacktriangleright^{N} v_j, v_j \triangleright v_k \vdash v_i \blacktriangleright^{N \cup \{k\}} v_k$

Chain $v_i \triangleright v_j, v_j \triangleright v_k \vdash v_i \diamondsuit^{\{j\}} v_k$

Fork $v_j \triangleright v_i, v_j \triangleright v_k \vdash v_i \diamondsuit^{\{j\}} v_k$

Collider $v_i \triangleright v_j, v_k \triangleright v_j, v_j \blacktriangleright^{N} v_z \vdash v_i \diamondsuit_{\{j,N\}} v_k$

Transitivity* $v_x \diamondsuit_N^M v_i, v_j \diamondsuit_P^O v_y \vdash v_x \diamondsuit_{N \cup P}^{M \cup O} v_y$

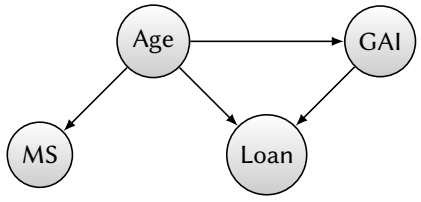
$$\frac{\triangleright_{\widehat{f}}, \sigma \Big|\sim\ t : \delta_p}{\triangleright_{\widehat{f}}, \sigma, a : \alpha \Big|\sim\ t : \delta_p}\ Weakening^* \tag{4}$$

With *Conditions* consisting of:

Condition1: $a \not\triangleright t$ and $t \not\triangleright a$;

Condition2: For every $a \diamondsuit_N^M t$ in $\diamondsuit_{\widehat{f}}$, either $M \cap Var_\sigma \neq \emptyset$ or $\exists S \in N(\exists x \in S \wedge S \cap Var_\sigma = \varnothing)$.

Hence, to decide whether Weakening can be applied, we check both $\triangleright_{\widehat{f}}$ and $\diamondsuit_{\widehat{f}}$ to evaluate conditional independence. In particular, the first condition requires that protected variable and target variable are not one the direct cause of the other, and the second condition requires that, for every path connecting them, σ blocks it, by containing at least one non-collider or by not containing a collider and all of its descendants. Note that this rule can be used to decide both **IF** in general and intersectionality, which corresponds to cases in which σ already contains a protected attribute. More precisely, what is obtained by checking the admissibility of an instance of Weakening is an evaluation of fairness (and possibly intersectionality) of the classifier, when a specific set of attributes are used to decide a target variable. Figure 3 shows a simple example of application of this rule.



(a) Causal graph of the classifier.

$$\frac{\triangleright_{\widehat{f}}, Age : 27,\ GAI : 40K \Big|\sim\ Loan : yes_{0.60}}{\triangleright_{\widehat{f}}, Age : 27,\ GAI : 40K,\ MS : m \Big|\sim\ Loan : yes_{0.60}}\ W^*$$

(b) Application of W with attribute MS receving value m (married). The satisfaction of the conditions can be observed from the causal graph, and is derivable from the set $\triangleright_{\widehat{f}}$.

Figure 3: Example of application of Weakening.

4. Conclusion

This work focuses on formal tools to check fairness of probabilistic classifiers. We have shown that, without taking into account the causal relations between the features of the classifier, intersectional fairness is not guaranteed. The proposed typed natural deduction calculus **TNDPQ** has labels representing causal relations, and it provides a criterion of applicability for the rule of Weakening that establishes both fairness and intersectionality. An extension using causal labels to express counterfactual fairness is left for further work.

Acknowledgments

This research was supported by the Ministero dell'Università e della Ricerca (MUR) through PRIN 2022 Project SMARTEST – Simulation of Probabilistic Systems for the Age of the Digital Twin (20223E8Y4X),

and through the Project “Departments of Excellence 2023-2027” awarded to the Department of Philosophy “Piero Martinetti” of the University of Milan.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: Grammar and spelling check. After using these tool, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] F. A. D’Asaro, F. A. Genco, G. Primiero, Checking trustworthiness of probabilistic computations in a typed natural deduction system, Journal of Logic and Computation (2025) exaf003. URL: https://doi.org/10.1093/logcom/exaf003. doi:10.1093/logcom/exaf003.

[2] E. Kubyshkina, G. Primiero, A possible worlds semantics for trustworthy non-deterministic computations, International Journal of Approximate Reasoning 172 (2024) 109212. URL: https://www.sciencedirect.com/science/article/pii/S0888613X24000999. doi:https://doi.org/10.1016/j.ijar.2024.109212.

[3] G. Coraglia, F. A. D’Asaro, F. A. Genco, D. Giannuzzi, D. Posillipo, G. Primiero, C. Quaggio, Brioxalkemy: a bias detecting tool, in: G. Boella, F. A. D’Asaro, A. Dyoub, L. Gorrieri, F. A. Lisi, C. Manganini, G. Primiero (Eds.), Proceedings of the 2nd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AI*IA 2023), Rome, Italy, November 6, 2023, volume 3615 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 44–60. URL: https://ceur-ws.org/Vol-3615/paper4.pdf.

[4] G. Coraglia, F. A. Genco, P. Piantadosi, E. Bagli, P. Giuffrida, D. Posillipo, G. Primiero, Evaluating ai fairness in credit scoring with the brio tool, 2024. URL: https://arxiv.org/abs/2406.03292. arXiv:2406.03292.

[5] L. Ceragioli, G. Primiero, Trustworthiness preservation by copies of machine learning systems (submitted). URL: https://doi.org/10.48550/arXiv.2506.05203.

[6] M. Kearns, S. Neel, A. Roth, Z. S. Wu, Preventing fairness gerrymandering: Auditing and learning for subgroup fairness, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 2564–2572. URL: https://proceedings.mlr.press/v80/kearns18a.html.

[7] M. Nielsen, E. Gissi, S. Heidari, R. Horton, K. Nadeau, D. Ngila, S. Noble, H. Paik, G. Tadesse, E. Zeng, J. Zou, L. Schiebinger, Intersectional analysis for science and technology, Nature 640 (2025) 329–337. doi:10.1038/s41586-025-08774-w.

[8] M. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual fairness, volume 30, Massachusetts Institute of Technology Press, 2017, pp. 4067–4077.

[9] N. Asher, L. De Lara, S. Paul, C. Russell, Counterfactual models for fair and adequate explanations, Machine Learning and Knowledge Extraction 4 (2022) 316–349. URL: https://www.mdpi.com/2504-4990/4/2/14. doi:10.3390/make4020014.

[10] J. Pearl, M. Glymour, N. Jewell, Causal Inference in Statistics: A Primer, Wiley, 2017. URL: https://books.google.it/books?id=L3G-CgAAQBAJ.

[11] Viganò, Labelled Non-Classical Logics, Springer, New York, 2000.

[12] S. Negri, J. von Plato, Proof Analysis: A Contribution to Hilbert’s Last Problem, Cambridge University Press, Cambridge, 2011.