Università di Milano Bicocca
Department of Informatics, Systems and Communication
Ph.D. Program in Computer Science
Cycle xxxvi

Fair Classification with Explicit Constraints: a Neuro-symbolic Approach with Logic Tensor Networks

a dissertation by Greta Greco
registration number 869354

Supervisor: Prof. Matteo Palmonari
Tutor: Prof. Raimondo Schettini

january 2024

Abstract

Algorithms are vulnerable to biases that might render their decisions unfair toward particular groups of individuals. Fairness comes with a range of facets that strongly depend on the application domain that consider different notions of what is a fair decision in situations impacting individuals in the population. The precise differences, implications and orthogonality between these notions have not yet been fully analyzed and we try to make some order out of this zoo of definitions. When it comes about enforcing such constraints, most in-processing mitigation models embed fairness constraints as fundamental component of the loss function thus requiring code-level adjustments to adapt to specific contexts and domains. Rather than relying on a procedural approach, we propose a model that leverages declarative structured knowledge to encode fairness requirements in the form of logic rules, capturing unambiguous and precise natural language statements. We present a neuro-symbolic integration approach based on Logic Tensor Networks that combines data-driven network-based learning with high-level logical knowledge, allowing to perform classification tasks while reducing discrimination. Experimental evidence shows that our model is capable of encoding diverse definitions of fairness, covering a good portion of group constraints. With performance reaching or outperforming state-of-the-art approaches, our algorithm proves a flexible framework to account for non-discrimination with tiny to no cost at all in terms of accuracy.

Contents

Listing of figures

If a thing exists, it exists in some amount; and if it exists in some amount, it can be measured.

E. L. Thorndike

0 Preface

Humans are unwilling to trust machines as we are disinclined to attribute any sense of feeling, or moral consciousness, to algorithms. The way Artificial Intelligence models are designed though, strive for generalising from examples and faithfully capture underlying patterns. If training data mirrors biases and inequalities inherent to the real world, then it is legitimate to expect that algorithms are keen on replicating such disparities. Twist of fate, the

misgiving towards AI has opened the unprecedented opportunity of inspecting discrimination in recorded data, as unfair machines follow just as consequence, mostly. Information that fuels models have proved to be often tainted with prejudices and stereotypes, intrinsically pervasive in most phenomena and often affecting human decisions. Stressing the value of non-discrimination becomes redundant here, being universally recognised as fundamental principle of human rights. Hence, the importance of fair decision should disregard the nature of its origin, yet there is a growing concern towards machine-based decision.

At the time of writing this thesis, I went to the post office to pick up a parcel. A young man, not older than a teenager, had patiently waited his turn to ask for assistance since his bank card was stuck inside the ATM. The employee retrieved the card and asked for identity papers. The boy showed his father's documents since the card was registered to his parent who, it will be clear in a moment, was probably thousands miles apart. The employee replied thoughtlessly that there was no way of getting the card back unless the account holder showed up in person within three days, otherwise the card would have been shredded. Afterwards, the procedure to get the money back would have turned to be extremely complex. The boy was in obvious difficulty since that card was probably everything he owned, but kindly asked some additional information in the attempt of getting the situation sorted. At the same time, an old lady, whose highest valuable task of the day probably consisted in getting fresh blueberries at the greengrocer's, started to complain loudly about the time she was wasting because of this young man's pointless insistence. This shouldn't be an important information, but the skin of this boy was black and I am ready to bet that things would have gone differently otherwise.

I had spent my entire life convinced that if something exists, then it can be measured. Yet

this interaction was not recorded nor evaluated, thus in all respect, it hadn't happened. I have wondered whether it was worth to write this piece of work, since this attention towards AI fairness probably rises from the fact that it is more convenient to blame a machine, rather than a person. Nonetheless, I am persuaded that issues around discrimination should not be uniquely sought into a wrong weight of a neural network or an incorrect leaf split of a random forest but rather in the greatest of evils: human brain.

1

Introduction

Ensuring just algorithmic decisions is directly tied to the more ambitious intent of protecting the rights of individuals and avoiding harm. As society becomes more attuned to issues of discrimination and justice, there is a major focus on ensuring that AI systems are designed and deployed in ways that are fair and equitable. Governments and regulatory

bodies around the world are increasingly requiring organizations to guarantee that their AI systems do not discriminate; businesses not paying attention to such constraints could face legal and reputational consequences.

The most harmful impacts of AI on individuals arise when opportunities are unjustly denied or resources are unfairly distributed as a consequence of algorithmic decisions. This work focuses on feature-based binary classification tasks, where one of the possible outcomes represents an unfavourable decision towards groups of individuals identified by a sensitive attribute, whose nature is typically socio- demographic. Binary classification represents *de facto* the archetypal task of Machine Learning, yet its ubiquity attests the key role it plays in the industry and, as far as I can witness, in the financial sector. This manuscript is the result of the research I conducted within an executive Ph.D program in collaboration with Intesa Sanpaolo, where I experienced that artificial Intelligence offers countless and impressive possibilities but nevertheless, what often drives the highest value boils down to a yes-no decision. Imagine a situation where a person asks for a loan, whose acceptance or refusal is determined by a machine learning model. Here, training instances are based on past human decisions that might have been prone to biases concerning gender or ethnicity for instance. Defence against this eventuality requires a set up of guidelines, processes and monitoring over the ML pipeline: during my concurrent experience as a data scientist in a major international bank, I've been in charge of developing and put into production a number of AI models, each time verifying a potential risk for discrimination. And each time, despite an apparent simplicity, the hardest task was assessing whether model outcomes could be considered *fair*. This task encompasses a landscape of fairness definitions that is nevertheless extremely rich and complex that had started to build up far before the advent of artificial intelligence, as we intend

it. To this extent, this manuscript to disentangle and analyze in depth the research question

> RQ1: WHAT ARE THE INTERDEPENDENCE AND RELATIONSHIPS AMONG EXISTING FAIRNESS DEFINITIONS?

The contribution, that was collected and published in *Castelnovo et al. "A clarification of the nuances in the fairness metrics landscape." Scientific Reports 12.1 (2022)*, lies not so much in the exhaustiveness of the taxonomy as in the attempt to clarify the nuances in the fairness metrics landscape with respect to the twofold dimensions — group vs. individual notions and observational vs. causality-based notions. We focus mainly on general and qualitative descriptions: building upon rigorous definitions we want to highlight incompatibilities and links between apparently different concepts of fairness.

Along with definitions, come their enforcing. The burst of interest has fueled the research: literature has proposed a number of bias mitigation strategies that may apply according to the specific fairness definition at stake, that we explore later in the same section. In doing so, we begin to frame where our model places within the broad category of mitigation strategies while identifying similar frameworks we can compare against. As will be apparent, most mitigation approaches expect fairness constraints to be embedded within the loss function, hindering the feasibility to make modifications and adapt to the specific domain's purpose. In turn, the option to express natural language fairness constraints into logical predicates allows for a wider leeway while addressing the task. The ability to embed declarative logical statements into neural models is one of the key features of neuro-symbolic integration (NSI) and introduces the chance to account for symbolic rules, in contrast to a mere procedural data-driven process. We illustrates the key features and properties of NSI, wrapping up the existing frameworks within a taxonomy, questioning if

RQ2 - THERE EXISTS AN APPROPRIATE NEURO-SYMBOLIC FRAMEWORK TO EMBED FAIRNESS CONSTRAINTS?

After identifying a plausible alternative, we explain the rationales behind the choice of Logic Tensor Networks and briefly summarise its components, features and functioning.

The main experimental contributions and the formal framing of fairness constraints into injected knowledge aims at attesting whether

RQ3 - CAN WE TRANSPOSE A STATISTICAL FAIRNESS CONSTRAINT INTO A FIRST-ORDER LOGIC AXIOM?

We focus on group fairness, and specifically, we start by taking into account the requirement imposed by statistical parity, which expects an equal predicted positive rate among groups identified by sensitive attributes. Our contributions rely on a first-order logic axiomatization of fairness constraint, along with a solid theoretical discussion about the choice of optimal fuzzy logic operators for connectives and quantifiers. Experimental setting is meant to asses

RQ4 - CAN LOGIC TENSOR NETWORKS CORRECTLY OPTIMISE FOR FAIRNESS WITHOUT EXCESSIVELY LOOSING ON CLASSIFICATION ACCURACY?

Evidence shows that our approach reaches results that often outperform similar models while providing the additional advantage of greater flexibility and ease of constraint declaration. The novelty of our approach, in fact, lies in that the user can specify fairness constraint in a unique logic predicate, whose impact on the model as a whole, can be incrementally controlled by a corresponding weight. The axiomatization and experimental results are summarised in *Greco et al. "Declarative Encoding of Fairness in Logic Tensor Networks.", Proceedings of* 26^{th} *European Conference on Artificial Intelligence (2023).*

The intuition behind this paper was extended and further developed in this manuscript, primarily with the intent of assessing to what extent

> RQ5 - CAN OUR APPROACH BE GENERALISED TO DIFFERENT NOTIONS OF FAIRNESS?

The demonstration encompasses the implementation of axioms based on the separation principle through tiniest adjustment, thus enhancing the value of FOL compositionality. While applying LTN in the fairness domain, we have observed a strong impact of the logical connectives and quantifier interpretation on the model efficacy and inference task and we propose a discussion that helps clarifying model choices that might benefit the model application in real-word use cases. Evidences illustrate that Logic Tensor Networks represent a valid approach in the wild and bungled context of fairness. Results proves as good as top state-of-the art models while providing extra avails. Undoubtedly, this work lays just a first brick but unlocks further research on neuro-symbolic-based bias mitigation strategies, for which we propose a few future directions.

The rest of this work is organised as follows: Section 2 introduces the main concept of algorithmic fairness in the attempt of clarifying their interrelations along with the most promising approaches for bias mitigation, which we compare against. Afterwards, we wrap up the background of neuro-symbolic integration approaches in Section 3 highlighting the key features. We explain the reason behind the choice of opting for Logic Tensor Networks in the attempt of enforcing group fairness in a binary classification task. The approach we propose is thoroughly described in Section 4 that collects the experiments and results concerning the fairness notion of independence. Last, we demonstrate that our model is capable of optimising for a diverse fairness notion based on separation in Section 5 where we extend the analysis on

implication interpretation and we extensively test details concerning the research question. We conclude with Section 6 where we summarise our findings, declare the limitations and collect future research directions.

2

Algorithmic Fairness

The discussion about fair decision making has only recently shifted to the level of Artificial Intelligence models. Despite its seemingly compelling nature, this debate has at least half-century of history (Hutchinson & Mitchell, 2019): its course has evolved trough a few decades in a way that partly mirrors and revisits the current trends. It all began in 1964 when

the landmark United States Civil Rights Act outlawed discrimination on the basis of race, color, religion, sex or national origin. The use of assessment tests in the public and private sector was an established practice at that time, particularly in the employment and educational systems, to begin with. Such tests generally consisted in simple linear models where the prediction was a (weighted) sum of item scores representing features of the participants. The very first attempts to quantitatively assess whether a model could be discriminatory were carried out by the psychologist and statistician T.Anne Cleary, researcher at the Educational Testing Services. She was concerned about regression models used for predicting educational outcomes from test scores (Cleary, 1966, 1968). According to her work, a test is said to be biased if consistent nonzero prediction errors are made for members of a (racial) subgroup. Namely, if the score predicted from the common regression line is consistently too low for members of a subgroup, the results are said to be unfair. During that same period, Robert Guion carried out analogous analysis on the employment domain, trying to assess whether candidates with similar probability of success had similar probability of being hired for the job. He uncovered the difficulties in objectively measuring certain features, like the probability of success for instance, which are not observable and often a sophisticated construct are needed for an estimation. A generalised concern and skepticism about evaluation tests started to emerge but the scientific community and civil rights associations agreed that interviews couldn't be a preferable alternative since indeed they could introduce an increasing amount of subjective bias.

In the seventies, another prominent figure that contributed to shape advancements in the field, Edward Thorndike, was convinced that the discussion had been oversimplified. He flips the view and formulates one of the first quantitative of fairness, rather than focusing

on *un*fairness. Differently from Cleary, that based her work on the notion of independence, (Thorndike, 1971) introduces a concept — known today as separation — that accounts for situations where different groups exhibit (legitimately) different base rates of the target variable. In this scenario, a test is said to be fair if the ratio of predicted positives to ground truth positives is equal for each group. He arguably pointed out two major issues that are yet again under continual discussion: the inevitable tension between the individual and group fairness concepts and the impossibility of decoupling the definition of fairness from the specific contexts it is used in. In other words, there is no such thing as a fair model, but a fair *use* of the model. In the attempt of finding a common formalization, Darlington (1971) expresses the two approaches in terms of correlation between the demographic attribute, the test score and the target variable, stressing the fact that the two are incompatible, except for under special conditions. Hence, a compromise situation could easily end up not satisfying either criterion. In the following years, fairness definition began to proliferate introducing the possibility to account for equality of true positive rates (Cole, 1973), equality of positive predictive values (Linn, 1973), equality of true negative and the ratio between true positive and true negative rates across subgroups (Petersen & Novick, 1976). Noteworthy, not all the authors that proposed a novel approach also advocated for its use: rather, they were explicitly exploring the space of possibilities (Hutchinson & Mitchell, 2019).

By the end of the decade, research came to a standstill. To date, the scientific community was unable to identify a statistic that could unambiguously indicate whether a test could be considered fair. The multitude of definition often at odds, the lack of clarity about their usage and the absence of a standard procedure kept any further advancements in check.

The complex and dynamic nature of the problem has reflected in a consistent response from the scientific community in the attempt to fill the gap between ethical regulatory demand and the increasing adoption of AI at scale. This work focuses on feature-based binary classification scenario, whose relevance has fostered extended research (Mehrabi et al., 2019) since in this circumstance fairness can be actually quantified by a number of different metrics (Castelnovo et al., 2022b) and eventually mitigated.

It is assumed the reader is familiar with classification tasks and we'll briefly introduce the problem for the sake of illustrating the notation that will be used throughout the manuscript. Let A be the categorical random variable representing a protected attribute, that for reference we shall take to be gender, we label with X all the other (non-sensitive) random variables that the algorithm is going to use to provide its yes/no decisions represented by $\hat{Y} = f(X, A) \in \{0, 1\}$. We label $Y \in \{0, 1\}$ the ground truth target variable that needs to be estimated – typically by minimizing some loss function $\mathcal{L}(Y, \hat{Y})$. $\widetilde{X} = (X, A)$ collectively represent all the features of the problem. We denote with lowercase letters the specific realizations of random variables, e.g. $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ represent a dataset of n independent realizations of (X, Y). We employ calligraphic symbols to refer to domain spaces, namely $\mathcal{X}$ denotes the space where features X live*.

An example of such task, that we will often exploit as a running example, consists in predicting whether a credit applicant will get a positive or negative response from an algorithm that considers his personal and financial features and that was trained on past examples made

*Technically, since we employ uppercase letters to denote random variables, it would be more proper to say that $\mathcal{X}$ is the image space of $X : \Omega \to \mathcal{X}$, where Ω is the event space, and that $x \in \mathcal{X}$. We shall nevertheless use this slight abuse of notation for sake of simplicity.

of officers' decisions. The predicted outcome can be inspected to assess whether the decision can be considered a *fair classification*, either from an *individual* or a *group* connotation (Dwork et al., 2012; Zemel et al., 2013).

2.2 Measuring inequalities

Fairness notions proposed in recent literature, and thus specifically conceived in reference to Artificial Intelligence models, are usually classified in broad areas: definitions based on parity of statistical metrics across groups identified by different values in protected attributes (e.g. male and female individuals, or people belonging to different ethicity); definitions focusing on preventing different treatment for individuals considered similar with respect to a specific task; definitions advocating the necessity of finding and employing causality among variables in order to really disentangle unfair impacts on decisions. These three broad classes can be further seen in light of two major distinctions (Castelnovo et al., 2022b):

- Observational vs. causality-based discriminates criteria that are based purely on observational distribution of the data from criteria that try to first unveil causal relationships among the variables at play in a specific situation (mainly through a mixture of domain knowledge and opportune inference techniques) and then assess fairness.

- Group (or statistical) vs. individual (or similarity-based) discriminates criteria that focus on equality of treatment among groups of people from criteria requiring equality of treatment among couples of similar individuals.

The contribution we propose, besides collecting the most prominent definitions, consists in analysing the relationship among the metrics and trying to put order in the fairness land-

scape, accounting for our first research question

> RQ1: WHAT ARE THE INTERDEPENDENCE AND RELATIONSHIPS AMONG EXISTING FAIRNESS DEFINITIONS?

2.2.1 Individual Fairness

The hunch behind individual fairness embeds the concept that any two individual who are similar with respect to a particular task should be classified similarly (Dwork et al., 2012). This is perhaps the most forthright conception of equality thar resemble the notion originally outlined by Aristotle (Binns, 2020), disregards any constructs and statistics. In other words, irrelevant differences between people should not lead to significant differences in their chance of a positive outcome in a classification model. This paradigm, as the name itself, bears as an alternative to group fairness as the latter is acknowledged to concern protections for groups rather than for individuals. Next, it offers the opportunity to forbid, or at least reduce, a number of discriminatory practices from explicit and implicit discrimination to redlining and tokenism (Dwork et al., 2012). It can detect and address the issue of cherry-picking random individuals from the unfavoured group with the mere intent of a selecting a proportionate number, disadvantaging in fact deserving applicants.

The above mentioned arguments have motivated the groundbreaking work of Dwork et al. (2012) who proposed a mathematical formalization, encompassing two pieces. The first is a similarity metric: a distance measuring the extent of similarity between two individuals x and y. The second is a function that measures the difference in the chances two individuals get as an outcomes of the decision model, intended as a mapping M from individuals to outcomes. Individual fairness requires that the distance between two individuals' outcomes is no greater

than their distance according to the similarity metric and is formalised by Lipschitz condition. Any two individuals x_i, x_j that are at distance $d(x_i, x_j) \in [0, 1]$ map to distributions $M(x_i)$ and $M(x_j)$, respectively, such that the statistical distance between the mappings is at most the distance between individuals.

$$D(M(x_i), M(x_j)) \leq d(x_i, x_j) \tag{2.1}$$

Distributions over outcomes observed by x and y are indistinguishable up to their distance $d(x, y)$. Resuming the running example on credit granting, Eq. 2.1 requires that any two individuals with similar assets, finances or job position shall be predicted with similar loan acceptance score.

As demonstrated in the original work, a convenient property of the Lipschitz condition lies in that, in special circumstance, it does imply statistical parity between certain subsets of the population, winning the goal of individual and group fairness at once.

The intuitive simplicity of the paradigm however, wryly collides with difficulties in its practical application (Fleisher, 2021). The similarity metric is task-specific and hard to generalise, thus requiring human insight and domain information on a case-by-case basis. It requires constant adaptation and careful consideration about task-relevant features, but it is also rather complicated, or even unrealistic (Zemel et al., 2013), to conceive a reasonable distance metric among individuals. Indeed, devising an additional step that involves human judgment introduces the possibility of further bias. Apart from the obstacles in its practical application, the construction of the initial mapping suffers from two main drawbacks. First, the task of fair classification appears to boil down to finding a convenient fair distance metric. For instance, one might decide that two individuals are similar if they have the same age, and

increasingly differ as their age gap grows. This difference of course can involve more than one variable in the hyperspace, and might coincide with the Euclidean distance. Secondly, the framework does not involve a learning process: the mapping is formulated on a given set of individuals and does not provide a procedure to generalise when dealing with novel data.

A parallel stream that points in the direction of individual fairness, defines similar individuals as couples belonging to different groups with respect to sensitive features but with the same values for all the other features. Historically, in fact, the naive approach to the problem has been to assert that the algorithm merely does not consider protected attributes such as gender or ethnicity. To wit, this option is hence requiring that model outcome should remain unchanged if we take an observation and we only change its protected attribute A. This concept is usually referred to as Fairness Through Unawareness (FTU) or blindness (Verma & Rubin, 2018), and it is expressed as the requirement of not explicitly employing protected attributes in making decisions [†]. Despite the appealing simplicity, this notion express the constraint in terms of how the model *works*, without providing adequate indication about *assessing* if the requirement is satisfied in the first place. Even so, there exists a couple of candidate metrics that might fit the role, the first being proposed by Zemel et al. (2013) and it estimates how much the decision $\hat{Y}$ is close to the decisions given to its k nearest neighbors in the feature space:

$$\text{consistency} = 1 - \frac{1}{n}\left(\sum_{i=1}^{n}\left|\hat{y}_i - \frac{1}{k}\sum_{x_j \in kNN(x_i)}\hat{y}_j\right|\right), \tag{2.2}$$

[†]This concept is also referred to as disparate treatment, i.e. there is disparate treatment whenever two individuals sharing the same values of non-sensitive features but differing on the sensitive ones are treated differently (Barocas & Selbst, 2016; Zafar et al., 2017).

Similarly, Thanh et al. (2011) suggest that, given a couple of individual, one can devise discrimination if there exists a significant difference of treatment among the neighbors belonging to a protected group and its neighbors not belonging to it. Nonetheless, especially in circumstance where individuals exhibiting a given attribute are underrepresented (e.g. female), it may happen that the k neighbors of, say, a male individual are all males: in this case consistency (2.2) would in fact be equal to 1, but this does not prevent the model from explicitly using A in making decisions.

A second suitable option draws from the work of Berk (2009) and measures the difference in decisions among men and women weighted by their similarity in feature space:

$$\frac{1}{n_1 n_2} \sum_{\substack{a_i=1,\\ a_j=0}} e^{-dist(x_i,x_j)} |\hat{y}_i - \hat{y}_j|, \tag{2.3}$$

The higher its value the higher the difference in treatment for couples of similar males and females. here, the term $e^{-dist(x_i,x_j)}$ can be interpreted with any measure of similarity of the points x_i and x_j.

As noticed thus far, and it will become evident further on, no fairness approach comes with no limitations, drawbacks or counterexamples. As regards FTU, it fails at taking into account the *redundant encoding*, namely the possible interdependence between the protected attribute A and the rest of the feature space X. There are almost always ways of predicting unknown protected attributes from other seemingly innocuous features, thus explicitly removing the sensitive attribute is not sufficient to remove its information from the dataset. To bring an example, in a dataset there might be a very low chance that a male and a female have similar values in all the (other) features, since gender is correlated with some of them,

income for instance. According to the specific domain and task under exam, one may or may not decide that (some of) these correlations are legitimate and do not represent a source of unfair discrimination. Section 2.2.3 will dwell more on this issue, specifically it will examine situation where there may be some information correlated to the sensitive attribute but still considered *fair*.

Broadly speaking, the Lipschitz notion expressed in eq. 2.1 is sometimes referred to as Fairness Through Awareness (FTA), as opposed to Fairness Through Unawareness: in fact, even if they share the same principle of treating equally similar individuals, FTA is generally meant to use similarity metrics that are problem and target specific, i.e. that derive from an *awareness* of the possible impact, while FTU is a simple recipe that does not depend on the actual scenario. All in all, no consistent novelties have been introduced thereafter, in terms of theoretical formalization of a definition for individual fairness. Conversely, a number of strategies involving learning have been proposed and implemented in the attempt of overcoming the main shortcomings of the original approach, and will be discussed in Section 2.3.

2.2.2 Group Fairness

Fairness notions expressed in terms of equalising statistical criteria over groups constitutes the dominant research approach in literature. The three main broad partitions of observational group fairness were wrapped up in the benchmark work of Barocas & Selbst (2016) and include *independence*, *separation* and *sufficiency*.

Independence

Independence is strictly linked to what is known as *Demographic Parity* or *Statistical Parity*, and states that the *decisions should be independent of any sensitive attribute*:

$$\hat{Y} \perp\!\!\!\perp A. \tag{2.4}$$

In case of a binary decision $\hat{Y} \in \{0, 1\}$, this constraint can be formalized by asking that

$$P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = b), \quad \forall a, b \in \mathcal{A}, \tag{2.5}$$

i.e. the ratio of loans granted to men should be equal to the ratio of loans granted to women. In other words, membership in a protected class should have no correlation with the decision. Although under a different name, Statistical Parity was first formulated in the work by Kamiran & Calders (2009) and embodies a concept that has become pervasive in the literature being particularly straightforward and with attractive mathematical properties. Figure 2.1 shows a very simple visualization of a model reaching demographic parity among men and women.

In order to actually measure the amount of disparity, it is common to use either maximum possible difference of positive prediction ratios (PPR): a difference close to 0 indicates a *fair* decision system with respect to A. Typically, some tolerance is considered by employing a threshold below which the decisions are still considered acceptable. As an alternative, especially when the base rate for the target label Y is very low — in anomaly detection settings for instance — one can seek for the minimum possible ratio of PPR among all couples of

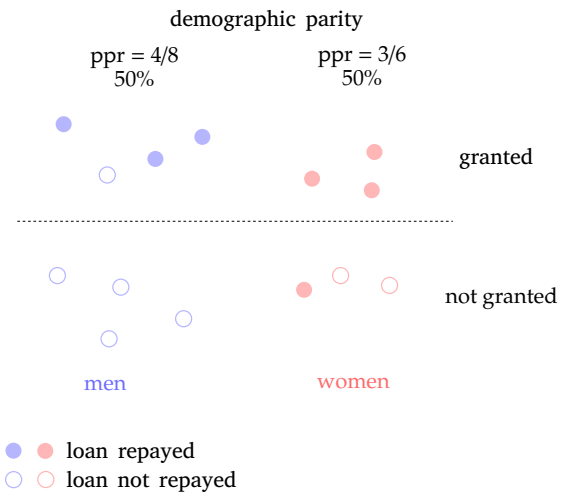


Figure 2.1: Example of demographic parity in gender in credit lending.

subgroups (being 1 its optimal value). This because in the said settings, even a remarkable difference in the predicted positive ratio among groups would result in a minimal and negligible difference to the untrained eye, while representing a big gap in all respects.

The intuition behind the concept is, however, merely superficial for a host of motivations. One of the counterexample that is often mentioned, concerns universal rejection: systematically denying the opportunity to all applicants would satisfy Statistical Parity while being patently unfair. This conjecture, which incidentally applies to many other criteria including individual fairness, ends in itself and strives for exploring the space of possibility, representing in fact a very unlike scenario. Additionally, fairness requirements in classification problems necessarily compete with the learning task, that in turn pursues optimization and the above mentioned scenario is rather unrealistic. A second, plausible and indeed perilous prospect concerns cases when, in order to satisfy independence, it is necessary to treat different groups in different ways. Precisely, in order to compensate for the (unwanted) model behavior, this requirement is likely to assign a favourable outcome to undeserving individual with the only

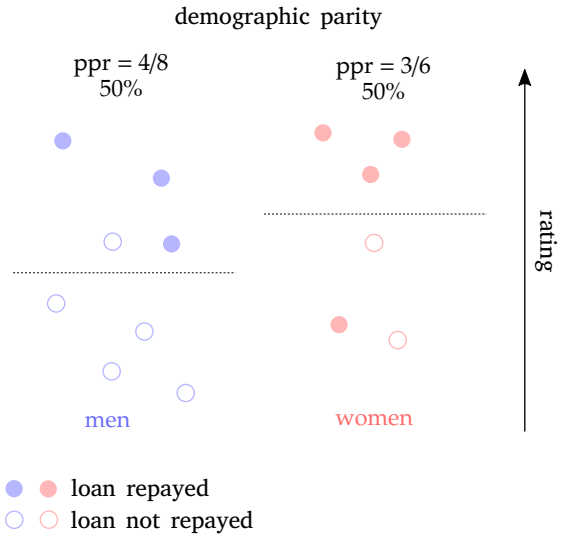


Figure 2.2: example of a subtlety of demographic parity: in order to reach demographic parity between men and women and still using rating as fundamental feature, one must use a different threshold between the groups, thus manifestly treating differently men and women.

goal of equalising the demographic leading to *treat different groups in different ways*. This is somehow the opposite of an intuitive notion of fairness. Let us refer to the context of credit lending and assume that, for whatever reason, women tend to actually pay back their loans with higher probability than men. If this is the case, it is reasonable to assume that a credit rating variable that we call R will be higher for women than for men. Figure 2.2 displays this example: to reach equal PPR, one must use a different rating threshold for each group.

Wrongly used, this gambit not only favours *reverse discrimination* denying opportunities to worthy candidates but can easily generate a vicious cycle in the longer term. Considering a loan application setting, granting a loan to unworthy individuals from the unfavoured groups, can lead to a situation where they might not be able to repay it back, lowering the overall credit worthiness of the said group, and in turn making it more difficult to get a loan in the future. Note that this last consideration does not necessarily arise on purpose: the

minority group might be underrepresented and the shortage of training data can result in a random guessing or even the model can overlook at prediction errors on such subgroup since they are less important in terms of mere counting. A last argument, that by no means exhaust the list, concerns the inter-relation between imposing independence and the impact on the classifier utility. Consider an oracle classifier that reaches perfect accuracy on a task where there is a legitimate correlation among A and $Y = \hat{Y}$, say in a context where the algorithm estimates the propensity to click on an advertisement. Here, it is reasonable to assume that independence shall not be verified at all cost and this isn't by itself a cause for concern as interests naturally vary from one group to another. As a result, the loss in utility of imposing demographic parity can be substantial for no good reason.

Despite its limitations, statistical parity represent a immediate and quite popular choice, would dare to say more frequently in literature than in industry application. For this reason, we'll start implementing this definition in our experiments illustrated in Section 4 both because we find a extensive ground for comparison with similar approaches and because, due to its simplicity, if our model cannot optimise on such a thing as statistical parity, it will hardly succeed on more complex definitions.

A variation of the independence criteria, that represent a reasonable requirement in many real-life scenarios, is *Conditional Demographic Parity*, first introduced in Kamiran et al. (2013). Instead of requiring full independence, one could condition on levels from another variable, let's say a risk levels. Formally, this results in requiring $\hat{Y}$ *independent of* A *given* R,

$$\hat{Y} \perp\!\!\!\perp A \mid R, \tag{2.6}$$

or, in other terms:

$$P(\hat{Y} = 1 \mid A = a, R = r) = P(\hat{Y} = 1 \mid A = b, R = r),$$
$$\forall a, b \in \mathcal{A}, \forall r. \quad (2.7)$$

In other words, the only disparities that are acceptable between predictions made on groups are those justified by a third, possibly *objective* and *reliable* variable. Resuming our running example, one could require equal acceptance rate of loan application *among* groups of applicants with the same credit rating. This goes somewhat in the direction of being an individual form of fairness requirement, since parity is assessed into smaller groups with respect to the entire sample. As a drawback, it might not be straightforward to identify the variables that are suitable for conditioning.

Separation

Independence and conditional independence over model outcomes, disregard the use of the true target Y. The concept of separation (Barocas & Selbst, 2016) instead, precisely conditions on the target label. This is equivalent to requiring the independence of the decision $\hat{Y}$ and gender A separately for individuals that actually repay their debt and for individuals that don't. Namely, among people that repay their debt (or don't), we want to have the same rate of loan granting for men and women. Formally, this concept can be expressed as follows:

$$\hat{Y} \perp\!\!\!\perp A \mid Y. \quad (2.8)$$

In other terms, placing the emphasis on equality among groups:

$$P(\hat{Y} = 1 \mid A = a, Y = y) = P(\hat{Y} = 1 \mid A = b, Y = y),$$
$$\forall a, b \in \mathcal{A},\ y \in \{0, 1\}. \quad (2.9)$$

Equivalently, disparities in groups with different values of A (male and female) should be completely justified by the value of Y (true repayment or not).

As in the conditional independence case, this appears a very reasonable fairness requirement, *provided that one can completely trust the target variable*. Namely, one should be extremely careful to check whether the target Y is not itself a source of bias.

For example, if Y instead of reflecting true repayment was the outcome of loan officers' decision on whether to grant the loans, it could incorporate bias, thus we it would be risky to assess fairness with direct comparisons with Y. Moreover, as said above, even in the objective case where Y is the actual repayment, a form of selection bias would likely distort the rate of repayment.

Separation can be expressed in terms of what are known in statistics as *type I* and *type II* errors. Indeed, it is easy to see that the two conditions in equation (2.9) — one for $y = 1$ and one for $y = 0$ — are equivalent to requiring that the model has the same false positive rate and false negative rate across groups identified by A. False positives and false negatives are precisely type I and type II errors, respectively. Namely, individuals that are granted loans but are not able to repay, and individuals that are able to repay but are not granted loans.

This is known as *Equality of Odds* (Hardt et al., 2016) and requires the same error rates across relevant group, as displayed in Figure 2.3.

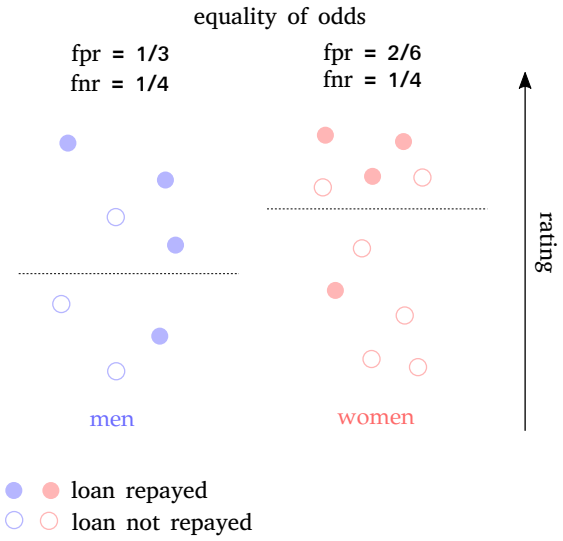


Figure 2.3: Example of Equality of Odds between men and women: false negative and false positive rates must be equal across groups.

. To encompass the great diversity of circumstances when it can be applied, there are two relaxed version of this criterion:

- *Predictive Equality*: equality of false positive rate across groups,

$$P(\hat{Y} = 1 \,|\, A = a, Y = 0) = P(\hat{Y} = 1 \,|\, A = b, Y = 0),$$

$$\forall a, b \in A,$$

- *Equality of Opportunity*: equality of false negative rate across groups,

$$P(\hat{Y} = 0 \,|\, A = a, Y = 1) = P(\hat{Y} = 0 \,|\, A = b, Y = 1),$$

$$\forall a, b \in \mathcal{A}.$$

While demographic parity, and independence in general, focuses on equality in terms of acceptance rate (loan grating rate), Equality of Odds, and separation in general, focuses on equality in terms of error rate: the model is fair provided that it is as efficient in one group as it is in the other.

The difference between Predictive Equality and Equality of Opportunity lies in the perspective behind the specific task. The first takes the perspective of people that won't repay the loan, while Equality of Opportunity takes the one of people that will repay. Depending on the problem at hand, one may consider either of these two alternatives as more important. For example, Predictive Equality may be considered when we want to minimize the risk of innocent people from being erroneously arrested: in this case it may be reasonable to focus on the parity of among innocents. In the credit lending environment, on the other hand, it may be reasonable to focus on Equality of Opportunity, i.e. on the parity among people that are indeed deserving. From the perspective of the banking industry, which correspond to the context of my executive program, our experiments reported in Section 5 will account for the notion of Equal Opportunity, demonstrating that it is possible to account for *one* of the separation notions but that could be indeed further extended to Predictive Equality, for instance,

To conclude, separation embodies the concept of parity given the ground truth outcome and see trough the eyes of individuals that are subject to the model decisions, rather than the decision maker. Conversely, the next subsection shall take into account the other side of the coin: parity given the model decision.

Sufficiency

Sufficiency (Barocas & Selbst, 2016) takes the perspective of people that are given the same model decision, and requires parity among them, irrespective of sensitive features. It takes into account the number of individuals who won't repay among those who are given the loan, differently from separation that deals with error rates in terms of fraction of errors over the ground truth — the number of individuals whose loan request is denied among those who would have repaid. Mathematically speaking, this is the same distinction that lies between recall (or true positive rate) and precision, namely $P(\hat{Y} = 1 \mid Y = 1)$ and $P(Y = 1 \mid \hat{Y} = 1)$, respectively.

A fairness criterion that focuses on this type of error rate is called *Predictive Parity* (Chouldechova, 2017), also referred to as *outcome test* (Verma & Rubin, 2018; Mitchell et al., 2021):

$$P(Y = 1 \mid A = a, \hat{Y} = 1) = P(Y = 1 \mid A = b, \hat{Y} = 1), \quad \forall a,\ b \in \mathcal{A}, \quad (2.10)$$

i.e. the model should have the same precision across sensitive groups. If we require condition (2.10) to hold for the case $Y = 0$ as well, then we get the following conditional independence statement:

$$Y \perp\!\!\!\perp A \mid \hat{Y},$$

which is precisely the condition required by sufficiency (Barocas & Selbst, 2016).

Predictive Parity, and its more general form of sufficiency, focuses on error parity among people who are given the same decision. In this regard, it takes the perspective of the decision

maker that tend to group people with respect to the decisions rather than the true outcomes. Taking the credit lending example, the decision maker is indeed more in control of sufficiency rather than separation, since parity given decision is something directly accessible, while parity given truth is known only in retrospect. Moreover, the group of people who are given the loan ($\hat{Y} = 1$) is less prone to selection bias than the group of people who repay the loan ($Y = 1$): indeed we can only have the information of repayment for the $\hat{Y} = 1$ group, but we know nothing about all the others ($\hat{Y} = 0$). As hinted at above, going along a similar reasoning, one can define other group metrics, such as Equality of Accuracy across groups: $P(\hat{Y} = Y | A = a) = P(\hat{Y} = Y | A = b)$, for all $a, b \in \mathcal{A}$, i.e. focusing on over unconditional errors, and others (Verma & Rubin, 2018).

2.2.3 Counterfactual Fairness

Another important distinction of fairness criteria occurs between *observational* and *causality-based* criteria.[‡] As we have seen, observational criteria rely only on observed realizations of the distribution of data and predictions. In fact, they focus on enforcing equal metrics (acceptance rate, error rate, etc...) for different groups of people. In this respect, they don't make further assumptions on the mechanism generating the data and suggest to assess fairness through statistical computation on observed data.

Causality-based criteria, on the other hand, try to leverage domain and expert knowledge to come up with a casual structure of the problem, that enables to answer questions like *what would have been the decision if that individual had a different gender?* While counterfactual questions seem in general closer to what one may intuitively think of as 'fairness assessment',

[‡]The modeling approach we propose for bias mitigation does not involve counterfactual fairness, therefore we shall only make a brief reference to cover the main aspects without delving into details.

the observational framework is on the one hand easier to assess and constrain on, and on the other more robust, since counterfactual criteria are subject to strong assumptions about the data and the underlying mechanism generating them, some of which are not even falsifiable.

As we argued above in section 2.2.1, answering to counterfactual questions is *very different* from taking the feature vector of, e.g., a male individual and just flip the gender label and see the consequences in the outcome. The difference lies precisely in the causal chain of 'events' that this flip would trigger. If there are some features related, e.g., to the length of the hair, or the height, then it is pretty obvious that the flip of gender should come together with a change in these two variables as well. And this may be the case for other, less obvious but more important, variables. This also suggests why counterfactual statements involve *causality relationships* among the variables. In general, to answer counterfactual questions, one needs to know the causal links underlying the problem. This requires a certain number of assumptions, usually driven by domain knowledge.

However, as major drawback, once given the casual structure there are many counterfactual models compatible with that structure (actually infinite), and the choice of one of them is not falsifiable in any way. Indeed, causality-based criteria can be formulated at least at two different levels, with increasing strength of assumptions: at the level of *interventions* and at the level of *counterfactuals* whose further discussion however, falls outside the scope of this manuscript.

2.2.4 Incompatibilities, conflicts and limitations

The proliferation of definitions has found little agreement on the most promising approach while none of them was proved to capture the intuitive notions of fairness (Fleisher, 2021).

What is more, there is a number of counterexamples suggesting that decisions satisfying statistical metrics over groups appear to behave blatantly unfair towards single individuals. To conclude, as pointed out by the well-known *impossibility theorem*, no more than one of the three fairness metrics of demographic parity, predictive parity and equalized odds can hold at the same time for a well calibrated classifier and a sensitive attribute capable of introducing machine bias (Kleinberg et al., 2016). The high count of metrics, definitions and paradigms that literature has proposed have not provided a contribution in the direction of clarifying when it is more appropriate to use which of the approaches. Remaining within the sphere of classification tasks, the main controversies that have arisen concerns three main arguments: what happens if one tries to satisfy multiple group criteria at once, whether it is possible to conciliate group and individual notions and how to tackle scenarios where more than a protected attribute shall be accounted for.

The impossibility theorem

While at first sight the major partition splits the individual and group paradigms, incompatibilities exists even, perhaps mainly, among different notions within group fairness. Undoubtedly, it is interesting to explore what happens if one requires to satisfy multiple criteria at once.

The short answer is that this is not possible *except* in trivial or degenerate scenarios, as stated by the following propositions drawn from the literature (Chouldechova, 2017; Kleinberg et al., 2016; Barocas et al., 2019; Räz, 2021; Simoiu et al., 2017):

1. *if Y is binary, $Y \not\perp\!\!\!\perp S$ and $Y \not\perp\!\!\!\perp A$, then* ***separation*** *and* ***independence*** *are incompatible*.

   To achieve both separation and independence, the only possibility is that either the

	notion		use of Y	condition
group fairness	Demographic Parity		-	equal acceptance rate across groups
	Conditional Demographic Parity		-*	equal acceptance rate across groups in any strata
	error parity	Equal Accuracy	✓	equal accuracy across groups
		Equality of Odds	✓	equal FPR and FNR across groups
		Predictive Parity	✓	equal precision across groups
individual fairness	FTU/Blindness		-	no explicit use of sensitive attributes
	Fairness Through Awareness		-*	similar people are given similar decisions
causality-based fairness	Counterfactual Fairness		-	an individual would have been given the same decision if she had had different values in sensitive attributes
	path-specific Counterfactual Fairness		-	same as above, but keeping fixed some specific attributes

* there are exceptions to these cases where Y is actually employed, e.g. CDP conditioning on Y becomes Equality of Odds, and there are notions of individual fairness that use a similarity metric defined on the target space (Berk et al., 2017).

Table 2.1: Fairness metrics: qualitative schema of the most important fairness metrics discussed throughout the manuscript.

model is completely useless ($Y \perp\!\!\!\perp S$), or the target is independent of the sensitive attribute ($Y \perp\!\!\!\perp A$), which implies an equal base rate for different sensitive groups. Namely, if there is an imbalance in groups identified by A, then one cannot have both conditions holding.

2. Analogously: *if $Y \not\perp\!\!\!\perp A$, then **sufficiency** and **independence** cannot hold simultaneously*.

   Thus, if there is an imbalance in base rates for groups identified by A, then you cannot impose both sufficiency and independence.

3. Finally, *if $Y \not\perp\!\!\!\perp A$ and the distribution (A, S, Y) is strictly positive, then **separation** and **sufficiency** are incompatible*.

   Meaning that separation and sufficiency can both hold either when there is no imbal-

ance in sensitive groups (i.e. the target is independent of sensitive attributes), or when the joint probability (A, S, Y) is degenerate, i.e. — for binary targets — when there are some values of A and S for which only $Y = 1$ (or $Y = 0$) holds, in other terms when the score exactly resolves the uncertainty in the target (as an example, the perfect classifier $S = Y$ always trivially satisfies both sufficiency and separation).

Notice that proposition 3 reduces to a more intuitive statement, if the classifier is also binary (e.g. when $S = \hat{Y}$): if $Y \not\perp\!\!\!\perp A$, Y and $\hat{Y}$ are binary, and there is at least one false positive prediction, then separation and sufficiency are incompatible. Moreover, it has be shown (Kleinberg et al., 2016) that Balance of Positive Class, Balance of Negative Class and Calibration within Groups can hold together only if either there is no imbalance in groups identified by A or if each individual is given a perfect prediction (i.e. $S \in \{0, 1\}$ everywhere).

When dealing with incompatibility of fairness metrics, the literature often focuses on the 2016 COMPAS[§] recidivism case (Angwin et al., 2016), that has become a case study in the fairness literature. Indeed, the debate on this case is a perfect example to highlight the fact that there are *different and non-compatible* notions of fairness, and that this may have concrete consequences on individuals. While literature has provided extensive discussion about the COMPAS case (Washington, 2018; Chouldechova, 2017), it is worth to point out that the debate involved two parties, one stating that the model predicting recidivism was *fair* since it satisfied Predictive Parity by ethnicity, while the other claiming it was *unfair* since it had different false positive and false negative rates for black and white individuals. Chouldechova (2017) showed that, if $Y \not\perp\!\!\!\perp A$, i.e. if the true recidivism rate is different for black and white people, then Predictive Parity and Equality of Odds cannot both hold. This suggests the

[§]A recidivism prediction instrument developed by Northpoint Inc.

foremost importance and careful reflection about which of the two notions (in general out of many others) is more important to be pursued in each specific case.

Summarizing, apart from trivial or peculiar scenarios, the three families of group criteria above presented are not mutually compatible. Recent works point in the direction of stating that two notions can be compatible provided that one does not seek perfect satisfaction of criteria. On the other side, this compromise might end up satisfying neither of the two.

On the conflict between individual and group fairness

One of the main drawbacks of group fairness lies in that it ensures conditions to hold only on average among groups, leaving room to bias discrimination *inside the groups*. Technically speaking, Statistical Parity can be satisfied using a different credit rating threshold for men and women leading to situations where, despite belonging to the same rating class, men will receive the loan, and women won't. This is precisely what individual fairness is intended to account for.

As already mentioned, this is only one possibility: one may as well reach DP by not using neither gender nor rating, and grant loan on the basis of other information, provided it is independent of A, but one would lose important information that can strongly penalise the classifier. Otherwise, in the attempt of pursuing DP by using as much information of Y as possible contained in (X, A) — thus minimizing the risk $E[\mathcal{L}(f(X, A), Y)]$ — then it is unavoidable to have some form of disparate treatment among people in different groups with respect to A whenever $X \not\perp\!\!\!\perp A$. This has been thoroughly discussed by Dwork et al. (2012), where they call *fair affirmative action* the process of requiring DP while trying to keep as low as possible the amount of disparate treatment between people having similar X.

In addition to what examined within the context of single definitions and in the attempt to clarify the general picture and help responding to RQ1, one can represent the different notions of (observational) fairness in a plane characterised by two *qualitative* dimensions (see figure 2.4): 1. to what extent a model is fair at the individual level, 2. how much information of A is retained in making decisions. The first dimension represents to what extent two individuals with similar overall features X are given similar decisions: the maximum value is reached by models blind on A (FTU). These are the models that are also using all information in X, irrespective of the interdependence with A, thus FTU-compliant models will use all information contained in $\widetilde{X}$ apart from the information that is contained in A only. The minimum value in this dimension is reached by models that satisfy DP. Models using suppression methods, being blind to both A and other features with high correlation with A, are individually fair in the sense of preventing disparate treatment. In so doing, they can exploit more or less information of A with respect to general DP-compliant models depending on how many correlated variables are discarded. However, the price to pay for discarding variables is in terms of errors in approximating Y, which is not highlighted in this plot. Notice that, of course, full suppression – i.e. removing all variables dependent on A – trivially satisfies the condition $\hat{Y} \perp\!\!\!\perp A$, i.e. it is DP-compliant as well. In other terms, one can have a DP-compliant model that is individual as well. In figure 2.4, we label with DP a general model that tries to maximize performance while satisfying a DP constraint, without any further consideration. Models satisfying CDP are somewhat in-between, of course depending on the specific variables considered for conditioning. They guarantee less disparate treatment than unconditional DP, and they use more information of A by controlling for other variables possibly dependent on A.

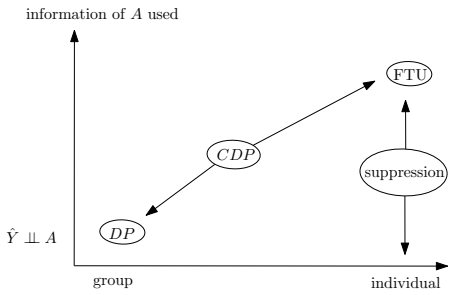


Figure 2.4: Landscape of observational fairness criteria with respect to the group-vs-individual dimension and the amount of information of A used (via X).

Notice that approaches such as fair representation (further discussed in Section 2.3.1), that try to remove all information of A from X to get new variables Z which are as close as possible to being independent of A, produce decision systems $\hat{Y} = f(Z)$ that are not, in general, individually fair. This is due to the simple fact that, precisely to remove the interdependence of A and X while keeping as much information of X as possible, two individuals with same X and different A will be mapped in two distinct points on $\mathcal{Z}$, thus having, in general, different outcomes. Referring again to the credit lending example, suppose that we have $R = g(A) + U$, with g a complicated function encoding the interdependence of rating and gender, and U some other factor independent of A representing other information in R *orthogonal* to A. In this setting, the variable Z that we are looking for is precisely U. Notice that U is indeed independent of A, thus any decision system $\hat{Y} = f(U)$ satisfies DP, but given two individuals with $R = r$ nothing prevents them from having different U. In other terms, you *need* to have some amount of disparate treatment to guarantee DP *and* employ as much information as possible to estimate Y.

Figure 2.5 shows a *qualitative* representation of observational metrics with respect to the amount of information of A (through X) that is used by the model, and the predictive per-

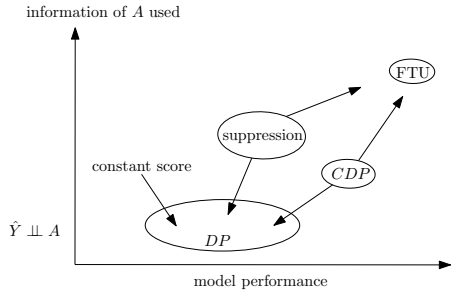


Figure 2.5: Landscape of observational fairness criteria with respect to the model performance dimension and the amount of information of A used (via X).

formance. Notice that DP can be reached in many ways: e.g. a constant score model, namely a model accepting with the same chance all the individuals irrespective of any feature, is DP-compliant (incidentally, it is also individual), a model in which all the variables dependent on A have been removed (a full suppression), or a model where DP is reached while trying to maximize performance (e.g. through fair representations). All these ways differ in terms of the overall performance of the DP-compliant model. FTU-compliant models, on the other hand, by employing all information in X will be, in general, more efficient in terms of model performance.[¶]

Notice that this discussion is to be taken at a qualitative level, one can come up with scenarios in which, e.g., models satisfying DP have higher performances than models FTU-compliant (think, e.g., of a situation in which $Y \perp\!\!\!\perp A$ and $X \not\perp\!\!\!\perp A$).

The (apparent) conflict that holds in principle between individual and group families of fairness notions is debated in the literature (Friedler et al., 2016; Hertweck et al., 2021; Binns, 2020). This analysis provides arguments to the claim that individual and group fairness are not in conflict *in general*, since they lie on the same line whose extreme values are separation

[¶]Of course it is understood that the model f in $\hat{Y} = f(X)$ is trained to maximize performance.

(i.e. unconditional independence) and FTU (i.e. individuals are given the same decisions on the basis of non-sensitive features only), with conditional metrics ranging between them. Thus, the crucial aspect is assuming and deciding *at the ethical and legal level* what are the variables that are *allowed* in specific scenarios: given those, we place ourselves on a point on this line reachable both by choosing an appropriate distance function on feature space (i.e. employing an individual concept) and by requiring parity among groups conditioned on those variables (a group notion).

Indeed, when referring to the conflict between individual and group families one is committing a slight *abuse of notation*, since the actual clash is rather on the *assumptions* regarding what is to be considered fair in a specific situation, than on metrics *per se*. Namely, one usually consider individual (group) concepts as the ones in which more (less) interdependence of X and A is allowed to be reflected in the final decisions. In the credit lending example, if income is correlated with gender, the issue weather it is fair to allow for a certain gender discrimination as long as justified by income can be seen as an instance of the conflict between individual or group notions, but it is rather a conflict about the underlying ethical and legal assumptions (Binns, 2020).

Coping with intersectional fairness and discretization

It is always not straightforward to identify which feature shall be treated as protected attribute and in fact, it is often the case where multiple variables must be considered jointly. However, all the definitions and results we gave in previous sections, at least in their plain formulations, posit that the sensitive feature A is represented by a single *and* categorical variable. Issues arise when any of the two conditions is no longer met.

Even though a detailed discussion about the issue of multiple sensitive features is out of the scope of this manuscript, we shall nevertheless give a brief overview. If, for a given problem, we identify more than one characteristic that needs to be taken into account as sensitive or protected – say $(A_1, \dots, A_l)$ – it is straightforward to assess fairness on each of them separately. This approach, that Yang et al. (2020) call *independent group fairness*, unfortunately is in general not enough: even if fairness is achieved (in whatever sense) separately on each sensitive variable A_i, it may happen that some subgroups given by the intersection of two or more A_i's undergo unfair discrimination with respect to the general population. This is sometimes referred to as *intersectional bias* (Crenshaw, 1994), or, more specifically, *fairness gerrymandering* (Kearns et al., 2018).

To prevent bias from occurring in *all* the possible subgroups identified by all A_i's one can simply identify a new feature $A = (A_1, \dots, A_l)$, whose values are the collection of values on all the sensitive attributes, and require fairness constraints on A. Yang et al. (2020) call this *intersectional group fairness*.

This last *trick* indeed solves the problem of intersectional bias, at least theoretically. Still, issues remain at a computational and practical level, whose two main reasons are:

- the exponential increase of the number of subgroups when adding sensitive features,
- the fact that, with finite samples, many of the subgroups will be empty or with very few observations.

These two aspects imply that assessing (group) fairness with respect to multiple sensitive attributes may be unfeasible in most practical cases. Since the presence of many sensitive features is more of a norm than it is an exception, this actually represents a huge problem, that

the literature on fairness in ML has barely begun to address (Yang et al., 2020; Kearns et al., 2018, 2019; Buolamwini & Gebru, 2018).

What has emerged so far, portray algorithmic fairness as rather chaotic and uncertain with very few firm points. This prevents regulations to provide precise indications about metrics or thresholds, leaving room for interpretations. What appears realistic while assessing and enforcing fairness, at least in the financial domain, can be summarised in a few rule of thumbs:

- Statistical parity, although immediate and easy to handle, it rarely reflects a plausible alternative in the business domain
- Notions base on error rate parity are indeed a more reasonable option. Here, instead of questioning about what is fair, it is helpful to identify what it *unfair*, namely what is the situation that lead to a possible discrimination, to best choose among Equal Opportunity and Predictive Parity
- It is often hard to assess group fairness when the minority group is heavily underrepresented and counts just a few instances. In those cases it is important to make sure the model does not learn to systematically overlook at those instances

There is no such thing as one-fits-all approach and this does not necessarily represent a negative aspect but implies the need to establish an apposite committee made up of different professionals, ranging from computer science to philosophy and legal, that can guide the business domain expert to come up with a plausible scenario.

2.3 Mitigating Disparities

Eliminating algorithmic bias is likely to represent a implausible intent, and this because of many reasons. A more reasonable objective points in the direction of mitigating and take control over model disparities, often setting up a tolerance threshold that again, is far from being universal. Loosely speaking, the existing means of mitigating algorithmic bias can be framed within three methods: pre-processing, in-processing, and post-processing, according to the step of the AI life cycle they operate on.

2.3.1 Acting on input data

Pre-processing is meant to reduce or eliminate bias in the dataset (Zhang et al., 2023) by *re-labelling* the target variable Y to satisfy a certain fairness measure (Kamiran & Calders, 2009; Luong et al., 2011), by *reweighting* (Calders et al., 2009; Krasanakis et al., 2018) or *resampling* (Kamiran et al., 2012) representative but unbiased instances or by *learning an intermediate representation* that satisfies fairness constraints while preserving helpful information (Zemel et al., 2013; Feldman et al., 2015). In some cases, the latter approach is often ascribed to a form of implicit in-processing since several approaches do involve learning, e.g. the use of variational autoencoder (Louizos et al., 2016) or adversarial learning Madras et al. (2018). All things considered, pre-processing techniques often lack the ability to achieve a user-defined trade-off between fairness and accuracy (Sun et al., 2022) and are not well suited to circumstances where the problem is caused by the algorithm. And besides, this category of approaches are often inconvenient to handle situations where the distributions of input data undergo frequent drifts.

2.3.2 Acting on model output

Post-processing strategies have the advantage of being model-agnostic and do not require access to the training procedure (Lohia et al., 2019). Loosely speaking, they function by changing the predicted labels on a subset of samples, appropriately selected to meet fairness constraints. Proposed methods differ in the rationales behind the choice of the instances that undergo a switch in the predicted labels: some approaches construct randomised decision rules (Hardt et al., 2016; Pleiss et al., 2017), others operate on the uncertainty boundary or the disagreement region of ensemble models (Kamiran et al., 2012) or imposing separate thresholds for different groups (Corbett-Davies et al., 2017; Menon & Williamson, 2018).

2.3.3 Acting at training time

If the mitigation is designed to be enforced at training time, it comes down to learning unbiased models on biased training data (Kamiran & Calders, 2009). Algorithms are then designed to maximise accuracy while minimising discrimination constraints. A prejudice remover regulariser has been proposed by Kamishima et al. (2011), which enforces a classifier's independence from sensitive information, to be integrated into the loss function of a logit model. Although the methods appears scalable to different classification models, the results in term of debiasing are only measured trough ad-hoc indexes, hindering to possibility to compare with other approaches. Donini et al. (2018) introduces an approach based on empirical risk minimization incorporating incorporates a fairness constraint into the learning problem. Another work that minimises the loss function subject to a series of constraints was presented by Zafar et al. (2017) that equalises misclassification rates. Remaining on learning with constraints, Goh et al. (2016) handles multiple goals on a single dataset using the

ramp penalty to accurately quantify costs. Severall fairness definitions are encompasses by Quadrianto & Sharmanska (2017) as a classical two-sample problem of conditional distributions, which can be solved using distance measures in Hilbert Space. Learning flexible representations that minimize the capability of an adversarial critic is introduced by Edwards & Storkey (2015). This adversary tries to predict the relevant sensitive variable from the representation, and so minimizing the performance of the adversary ensures there is little or no information in the representation about the sensitive variable. A fair neural network classifier (FNNC) was introduced by Padala & Gujar (2020) and incorporates fairness constraints into the loss in the form of Lagrangian multipliers. This work represent one valid method, as it covers different fairness definitions and reaches results that compare to state of the art and, most interestingly, is probably the most similar approach to what we propose. First, the two models share the neural backbone, relying a similar architecture made of a two-layered feed-forward neural network. Secondly, the fairness metric to be optimised, can be selected by acting on the model constraints. Nonetheless they differ in two yet crucial aspects: FNNC need to act on the loss function, hindering the possibility to further generalise to other definitions unless on has a profound control of code-level adjustments and moreover, the paper does not account for the possibility of tuning the relevance of the constraint in the computation of the overall task. This means that the model will try to reach the best possible debiasing, no matter the cost in terms of accuracy. Leveraging on a novel measure of decision boundary (un)fairness, Zafar et al. (2019) implemented two complementary approaches that maximise Disparate Impact and accuracy, respectively. Taking into account tree-based classifiers, Castelnovo et al. (2022a) presents FFTree, a method to find fair splits designed to work with different criteria and metrics.

Although characterised by different connotations, all the above-mentioned approaches rely on the idea of hard-coding the fairness notion as a building block of the loss function. Nonetheless, the need to ensure fair outcomes in a diversity of application domains requires a meticulous choice among the broad availability of definitions, or even the urge to devise a bespoke constraint. This makes it difficult to conceive and implement models able to respond to the most diverse requirements unless it is possible to leverage declarative knowledge, rather than procedural. When analysing results in Section 4 and 5, our approach will be experimentally compared against feature-based (binary) classification models that account for group fairness at training time, or that involve learning while retrieving a fair representation of the dataset. Specifically, we will take into account the major and most promising approaches that optimise Statistical Parity Difference, Disparate Impact or Equal Opportunity, as summarised in Table 2.2.

Finally, to the best of our knowledge, literature proposes a single work in the direction of neuro-symbolic integration for fairness (Wagner & d'Avila Garcez, 2021) that primarily insists on iterative querying to inspect biases through Shapely values and proposes interactive continual learning by adding knowledge through LTN. Conversely, this paper precisely focuses on fairness enforcement in binary classification tasks through first-order logic clauses instilled through LTN. We discuss the relationship with this previous work in Section 3.5.3.

2.3.4 Seeking individual parity

Despite this work focuses on bias mitigation in its group connotation, it is worth to mention that literature proposes a number of works that pursue individual fairness. Explicitly inspired by the theoretical background lied by the work from Dwork et al. (2012), Joseph et

Work	Type	Independence	Separation
Our Framework	In	DP, DI	EOpp
Padala & Gujar (2020)	In	DP, DI	
Castelnovo et al. (2022b)	In	DP, DI	EOpp
Madras et al. (2018)	Pre/In	DP	EOpp
Zemel et al. (2013)	Pre/In	DP	
Zafar et al. (2019)	In	DI	
Zafar et al. (2017)	In		EOpp
Feldman et al. (2015)	Pre	DI	
Hardt et al. (2016)	Post		EOpp
Donini et al. (2018)	Pre/In		EOpp
Edwards & Storkey (2015)	Pre/In	DP	
Goh et al. (2016)	In	DI	
Quadrianto & Sharmanska (2017)	In		EOpp

Table 2.2: Capabilities of different methods in mitigating disparities. Our experiments optimise DP and DI concerning the Independence notion and Equal Opportunity as a relaxed separation notions. We specifying which approaches account for different notions as a indicator of their ability to generalise. If not explicitly reported, specific metrics adopted in the original papers, differs from the ones we compute, therefore an immediate result confrontation on equal terms is not feasible.

al. (Joseph et al., 2016)propose a study based on the multi-armed bandit problem, interpreting non discrimination through demanding that given a set of applicants (say, for credit), a worse applicant is never favored over a better one, despite a model's uncertainty about the true payoffs. Their definition of fairness is analogous yet trying to overcome the assumption - and limitation - that the similarity metric is meant to be defined by the algorithm designer: the expected reward of each arm represent in fact a natural metric.

The problem with logic is that once you deduce something
you can't get rid of it.

M. Minsky

3

Learning with constraints: neuro-symbolic artificial intelligence

It should be no surprise that a *robust* artificial intelligence is to bear from a profound understanding of human mind, in terms of its functioning *and* ability to reason. For quite

a long time, artificial intelligence has been associated, as a matter of facts, with artificial neural networks, or connectionist AI, taken to be an abstraction of the physical workings of the brain (Hitzler et al., 2022). By contrast, what we perceive through introspection, and explicit cognitive reasoning can be understood as an abstraction of formal logic, namely symbolic AI. That being said, the early discussions about how to instill common sense into computers were a fair fight, all in all. In a Science article published in 1982 (Kolata, 1982), John McCarty, who also happened to be an inventor of the term *artificial intelligence*, was convinced that machines should be programmed to reason according to the well worked out language of mathematical logic, whether or not that was actually the way people think. The same article reported the opposite (today we would say complementary) view: putting large collections of information into a computer — much more information that is ever needed to solve a particular problem — and then define, in each situation, which details are crucial and which are optional. That time onward, the story is rather familiar: the emphasis on end-to-end learning with massive training sets has distracted from the more ambitious task of developing techniques involving high-level cognition (Marcus, 2020). The connectionist approach, however, has not revealed to be all sunshine and rainbows: concerns about trust, safety, interpretability and semantical soundness of AI are increasingly pervasive (Garcez & Lamb, 2020), matter if they come from the scientific community or regulatory bodies. Even discarding those risks, recent language model capabilities appear astonishing but not yet close to a *general* intelligence. It is still missing a sparkle that, according to many, is to come from explicit reasoning, whether it is instilled or emergent from a neural architecture. It's not like research hasn't focused on the integration of the neural and symbolic AI paradigms — the next sections will prove quite the opposite — but recently, more so than ever before, the need

for a principled synthesis has become evident.

Each paradigm complements each other with respect to their strengths and weaknesses. Deep learning and inference under uncertainty are expected to address the brittleness and computational complexity of symbolic systems. Learning from raw data, they are robust against outliers or errors in the underlying distributions, while symbolic systems are generally less trainable (Hitzler et al., 2022). Conversely, symbolic systems make explicit use of knowledge using formal languages — including logic — and the manipulation of symbols. They are to a high extent self-explanatory, as they can often be inspected and understood in detail by a human, aspect that isn't exactly connectionist flagship feature. Symbolism has been expected to provide additional knowledge in the form of constraints for learning, in the attempt of overcoming the neural difficulty with extrapolation in unbounded domains or with out-of-distribution data (Garcez & Lamb, 2020). However, despite several decades of research and a number of valid approaches, their mutual integration still remains a challenge.

3.1 Symbols and connections

Logical knowledge representation is symbolic in nature (Bader et al., 2004): the data structures under consideration consist of words over some language or of collections of finite trees, for example, depending on the perspective of the problem at hand. Logic programs, more specifically, consist of sets of first-order formulae under a restricted syntax and precisely, they are composed of a set of (universally quantified) clauses, which in turn consist of atoms and negated atoms only.

Successful connectionist architectures, however, can be understood as networks of simple computational units, in which activation is propagated as a real number. Input and output

of such networks consist of real valued vectors in Euclidean space in contrast to logic which is symbolic and thus discrete. Integrating logic and connectionism requires to bridge the gap between the discrete, symbolic setting of logic, and the continuous, real-valued setting of artificial neural networks.

3.2 What kind of integration

There are two driving forces behind the field of neural-symbolic integration: If on the one hand, the integration is the striving for an understanding of human cognition, on the other, the vision of combining the two technologies pursues more powerful reasoning and learning systems for computer science applications (Bader & Hitzler, 2005). The blending revolves around several aspects, some of them being rather technical, other more conceptual and cognition-oriented.

3.2.1 Localist versus Distributed

There is a bit of an issue around whether objects and concepts are represented in the brain by single neurons or multiple ones, where the multiple ones are supposed to represent subconcepts or microfeatures. Almost by osmosis, this conundrum has transferred to artificial intelligence, spawning the never ending debate between the theories of localist (in a sense symbolic) and distributed representation, generally associated with connectionism (Roy, 2011).

Symbolic representation of AI expects a single node in a network to represent a single entity, in contrast to the distributed representation, where such objects are represented in a decomposed form by their component elements or microfeatures. This decomposition principle works pretty well for physical objects but generally breaks down for higher-level con-

cepts in a semantic tree: modeling hierarchical semantic knowledge trough a connectionist approach tend to flatten the upper portion of the tree and represent the higher-level concepts by single nodes in the output layer of the network. This reduction results in a inevitable loss of semantic information. A second attention point concerns the distributions in microfeatures: although it appears to be rather an easy task to decompose single objects, a cat, into features like whiskers and tail, the corresponding higher-level concepts — such as a living thing and animal — are not substituted by any microfeature. These higher-level concepts are still represented by single nodes given the impossibility to be represented in terms of any physical or perceptual attributes (as the ones used for physical objects).

The success of connectionism suggests that distributed representations with gradient-based methods are more adequate for learning and optimization. However, extensive literature (Rogers & McClelland, 2004; McClelland & Rogers, 2003) has been unable to demonstrate that semantic relations can be learned without using the corresponding symbolic representations of concepts that therefore needs to be an inherent and necessary part of neural models of hierarchical semantic knowledge. Thus the need for single node symbolic representation of higher-level concepts is clearly evident in these models of complex cognitive processes.

3.2.2 Integrated versus Hybrid

The interrelations among the two paradigms can occur through an hybrid system, meaning that to address a problem, one can combine the two solving techniques that run in parallel. An integrated neural-symbolic system differs in that it consists of one connectionist main component in which symbolic knowledge is processed. A detailed taxonomy of said approaches is illustrated in Section 3.4

3.3 What kind of Language

The knowledge representation language that systems are able to deal with embodies a key capability of logic-based neural-symbolic integration approaches. In this respect, a major distinction needs to be made between systems based on propositional logics, and those relying on first-order predicate logics(Bader & Hitzler, 2005).

3.3.1 Propositional versus First-Order

Propositional theories involve only a finite number of propositional variables, and corresponding models. They do not require sophisticated symbol processing to handle nested terms, as well as substitutions or unifications. Due to their finitary nature and a corresponding ease of implementation, propositional logic programs using neural networks has represented the major line of research. To begin with, the landmark work by McCulloch & Pitts (1943) provides fundamental insights on how propositional logic can be processed using simple artificial neural networks.

In contrast, predicate logics generally allow to use function symbols as language primitives introducing the possibility to use terms of arbitrary depth. Models shall necessarily assign truth values to an infinite number of ground atoms. The difficulty lies in the finiteness of neural networks, that yet necessitates to capture the infinitary aspects of predicate logics by finite means(Bader & Hitzler, 2005). First-order approaches attempt to tackle the issue by using encodings of infinite sets by real numbers, and representing them in an approximate manner. Another problem is linked to variable binding, and refers to the fact that the same variable may occur in several places in a formula, or that during a reasoning process variables may be bound to instantiate terms. In a neural setting, different parts of formulae and differ-

ent individuals or terms are usually represented independently thus subnets are blind with respect to detailed activation patterns in other regions but this issue has been dealt with by a number of different expedients. That being said, the longstanding limitation that allowed to embed just fragments of first-order logic, has been recently unlocked (Badreddine et al., 2022) in the direction of full first-order logic, offering incredibly numerous attributes, including expressiveness, symbolic manipulation of variables, compositionality and modularity to begin with. This was made possible by translating logical statements into the loss function rather than into the network architecture. First-order logic statements are therefore mapped onto differentiable real-valued constraints using a many-valued logic interpretation. The trained network and the logic become communicating modules of a hybrid system, instead of the logic computation being implemented by the network.

3.4 Integration Taxonomies

It appears to be a natural question to ask how these two abstractions can be related or even unified in a principled way, or how symbol manipulation can arise from a neural substrate. Statistical learning and symbolic reasoning have been developed largely by distinct research (D'Avila Garcez et al., 2015) and one of the first attempts to summarise different approaches in a taxonomy was proposed by Henry Kautz in his talk with AAAI Conference in 2020 (Wang & Yang, 2022).

- Type 1. Symbolic Neuro Symbolic: this is the current standard operating procedure of deep learning methods in some application tasks where the input and output are symbols. Includes many Natural Language Processing models where the symbolic input is converted into vector representations, or embeddings. Real valued vectors are

then processes by the network that, at the end, transfers them to the required symbolic category or sequence of symbols via a softmax operation.

- Type 2. Symbolic[Neuro]: includes loosely coupled hybrid systems, where neural modules are internally used as subroutine within a symbolic problem solver.

- Type 3. Neuro | Symbolic: differs from type 2. in that the neural component is a parallel complementary program rather than of a sub-routine, and focus on a collaboration instead to a dependence.

- Type 4. Neuro:Symbolic $\rightarrow$ Neuro: Symbolic knowledge is compiled into the training set of a neural network. This includes tightly-coupled but localist neural-symbolic systems, where various forms of symbolic knowledge, not restricted to if-then rules, are translated into the initial architecture and set of weights of a neural network. Differently from Type 1., this does not provide a symbolic derivation of the result.

- Type 5. $\text{Neuro}_{\text{Symbolic}}$: tightly-coupled but distributed neural-symbolic systems, with a direct encoding of logical statements into neural structures through the use of embeddings, acting as a form of regularization and soft constraints on the network loss function. Logic Tensor Networks falls in this category as it translates first-order logic formulae as fuzzy relations on real numbers for neural computing, allowing gradient based sub-symbolic learning.

- Type 6. Neuro[Symbolic]: inspired by System1 and System2 described by Kahneman (2017), is a fully-integrated system capable of true symbolic reasoning inside

a neural engine where. System1 (neural) conducts initial reasoning, and when it puts attention on a certain part of the problem, it triggers System2 (symbolic) which performs a combinatorial search. It can perform challenging tasks that requires the ability to efficiently perform complex symbolic reasoning and search through large solution spaces.

3.5 Choosing an appropriate framework for fairness

3.5.1 Rationales behind the choice

Among different approaches provided by literature, for the purpose of enforcing group fairness through constraints injection, we evaluated Logic Tensor Networks to be a suitable framework, thus tackling

> RQ2 - there exists an appropriate neuro-symbolic framework to embed fairness constraints?

The reasons behind this choice are multiple:

- Expressiveness: based on differentiable first-order logic with fuzzy semantics, it supports different interpretations of logical connectives and quantifiers, which enables the use of the full expressiveness of FOL and several modeling choices
- Compositionality/Modularity: allows to refer to larger or nested clauses by the composition of symbols and relations among them. This ingredient offers the most promising advantages in the attempt of encoding fairness since, in principle, can allow to take into account different notions of group fairness, including Independence and separation, that are analysed within this work

- Clause weighting: LTNs allow for the specification of weights on clauses, which can be highly beneficial when dealing with conflicting tasks. This capability is particularly useful in balancing competing objectives and in our domain, to retain control over the desirable fairness level

- Versatility: LTNs allow to mix predicates whose interpretation is learned from the data and predicates whose interpretation is fixed into individual formulas, which we found relevant to model fairness

- Undirected graphical models: views logic as a constraint on a predictive model rather than focusing on causal relationships. This aligns better with the notion of fairness being viewed as a constraint on the model

3.5.2 Logic Tensor Networks

Logic Tensor Networks (Badreddine et al., 2022) are a NSI framework and computational model that enables learning and reasoning using rich knowledge, that relies on Real Logic (Serafini & d'Avila Garcez, 2016), way more expressive than propositional logic. LTN supports the representation and computation of the main AI tasks, including binary classification, leveraging on a uniform language. Knowledge can be expressed through fully differentiable first-order logic, encompassing universal and existential quantification (such as $\forall x$ and $\exists y$), arbitrary n-ary relations over variables (for example, $R(x, y, z, ...)$), and function symbols.

As mentioned, Real logic constitutes a major component of the framework and can be summarized through its most distinguishing features.

- **Syntax**: Real Logic is defined on a first-order language $\mathcal{L}$ with a signature that contains a set $\mathcal{C}$ of constant symbols, a set $\mathcal{F}$ of function symbols, a set $\mathcal{P}$ of relation symbols (predicates), and a set $\mathcal{X}$ of variable symbols. Terms are constants and variables (objects), *sequences of terms*, and function symbols applied to terms. *Objects, functions, and predicates are typed*: a function **D** assigns types to the elements of $\mathcal{L}$ to the corresponding domain symbol $\mathcal{D}$. Formulas are defined as in FOL, provided that typing constraints are preserved: if t_1,t_2, and t are terms and P is a predicate, then $t_1 = t_2$ and $P(t)$ are atomic formulas; complex formulas can be defined inductively as usual with logical connectives.

- **Semantics** is inspired by standard abstract semantics of FOL and based on a *grounding function* $\mathcal{G}$, which provides the interpretation of the domain symbols in $\mathcal{D}$ and the non-logical symbols in $\mathcal{L}$. $\mathcal{G}$ associates a tensor of real numbers to any term of $\mathcal{L}$, and a real number in the interval $[0, 1]$ to any formula φ. Intuitively, $G(t)$ are the numeric features of the objects denoted by t, and $G(\varphi)$ represents the system's degree of confidence in the truth of φ; the higher the value, the higher the confidence. There are a few aspects of LTN semantics that are relevant to our work. The tensor operation that grounds a predicate P can be implemented with an arbitrary neural network; the semantics of logical connectives is based on the semantics of first-order fuzzy logic, for which different interpretations have been proposed (e.g., different t-norms and t-conorms for conjunction and disjunction), thus making LTN highly dependent on the selection of specific semantic interpretations; if domain knowledge, in our case fairness, highly depends on universally quantified implications, the interpretation of fuzzy implication have a deep impact on modeling (e.g., Gödel vs Łukasiewicz); differ-

ent parametric interpretations of the universal quantifiers are possible: the truth value of a formula φ such as $\forall x A(x)$ (where $A(x)$ represents an arbitrary formula with a free variable x) estimates the truth value of φ based on some aggregation of the truth values estimated for instantiated formulas $A(c)$ found in the training data.

- **Learning** is mainly governed by grounding that plays a key role in the task of inductive inference. After fixing some choices, e.g., interpretation of connectives, interpretation of quantifiers, and boundaries of domain grounding, the parameters that underpin the representations of language elements can be learned in such a way as to maximize the satisfiability of a set of axioms, which include factual propositions available in a training set as well as generalized propositions encoding general constraints. The learning process eventually searches the optimal set of parameters from a hypothesis space that maximises the satisfiability of a theory $\mathcal{T} = \langle \mathcal{K}, \mathcal{G}_\theta \rangle$ namely the tuple composed by the set of closed predicates $\mathcal{K}$ defined on a set of symbols, and the parametric grounding for symbols and logical operators $\mathcal{G}_\theta$.

Two powerful concepts of LTN are (1) the grounding of logical concepts onto tensors with the use of logical statements which act as constraints on the vector space to help learning of an adequate embedding, and (2) the modular and differentiable organisation of knowledge within the neural network which allows querying and interaction with the system. Any user-defined statement in first-order logic can be queried in LTN which checks if that knowledge is satisfied by the trained neural network.

3.5.3 Previous literature

Given the premises, if natural language facts or constraints, can be formalised into logical statements, then fairness can be instilled through a neural-symbolic approach and, interestingly enough, little research has been conducted in this direction. As mentioned previously, a first approach to neural-symbolic integration to fairness was proposed by (Wagner & d'Avila Garcez, 2021) that, differently from our objectives, primarily focuses on the continuous interaction between the model and the human in the loop. Using explanatory features provided by Shapeley values, the authors recursively inspect bias in model outcomes and address it accordingly by injecting background knowledge. This work is based on a previous version of LTN that is no longer available and that differs from the current version in a number of relevant theoretical and code-level aspects, and does not investigate the role of different axioms and interpretations of logical operators. By contrast, with our work, we propose an original axiomatization of fairness that, to the best of our knowledge, is unseen in literature, and provide a meticulous investigation of the role played by different mathematical interpretations of universal quantifiers and implication operators.

4

Harnessing disparities with Logic Tensor Networks

Further on we'll try to answer our research questions regarding the feasibility of seeking group fairness trough a first-order logic axiomatization of constraints to be used within the

neuro-symbolic framework of Logic Tensor Networks. We'll build the knowledge base and test different choices of quantifiers' parameters, implication interpretation and the importance of fairness task within the overall optimisation

4.1 Problem setting

Framing a classification task within LTN requires the definition of the knowledge base along with the encoding of dataset features and classification labels using Real Logic. To begin with, let `Instances` denote the **domain** of the examples in the dataset. If the training examples are described by n features, then the grounding of the domain can be expressed as

$$\mathcal{G}(\mathit{Instances}) \subseteq \mathbb{R}^n \tag{4.1}$$

Therefore, each example k of domain `Instances` is a **constant** symbol whose grounding $\mathcal{G}(k)$ is represented by a tensor in $\mathcal{G}(\mathbf{D}(k)) = \mathbb{R}^n$. We can then introduce a **variable** x that represents a finite sequence of m individuals, each described by n features:

$$\mathcal{G}(x) = \mathbb{R}^{m \times n} \tag{4.2}$$

In the credit granting context, instances represent individuals to be classified, described by their features (e.g. income or age).

To retrieve information about the target variable, we introduce the function *Label* that maps each instance k to its corresponding label $y \in 0, 1$. This helps in identifying two variables x_+ and x_- indicating respectively the sequence of positive and negative training instances. The grounding of the variable indicating positive instances can be expressed as

follows:

$$\mathcal{G}(x_+) = \langle d \in \mathcal{G}(x) \mid \mathit{Label}(d) = 1 \rangle \tag{4.3}$$

Finally, we introduce a set of **predicates** that operate on the domain and represent classes of these individuals. Predicates can be known a priori, for instance they can identify whether an instance belongs to the privileged - or unprivileged - subgroup: for instance $\mathit{Priv}(x)$ and $\mathit{Unpriv}(x)$ are non-trainable predicates that map each individual x to a truth value based on whether he belongs to the privileged or non-privileged group, respectively. Alternatively, predicates can be predicted – and are thus trainable – as it happens for the classification task that in our case predicts learns and predicts whether an individual shall be granted or denied a loan. The classification task is thus accomplished by predicate $\underline{C(x)}$ [*], a trainable classifier with grounding

$$\mathcal{G}(\underline{C} \mid \theta) : x \rightarrow \texttt{sigmoid}(\texttt{MLP}_\theta(x)) \tag{4.4}$$

where MLP is a Multilayer Perceptron [†] with learning parameters θ and single output neuron that returns values between 0 and 1, interpreted as truth values. The learning process is meant to optimise the parameters of the predicate functions $\underline{C(x)}$ while satisfying the following constraints:

$$\begin{gathered} \forall x_+ \ \underline{C(x_+)} \\ \forall x_- \neg \underline{C(x_-)} \end{gathered} \tag{4.5}$$

The grounding of instances remains stable and is not subject to training. This possibility is encompassed by the framework we use and could be explored in the future; for the intent of this work however, we focus on a pure in-processing technique without learning a (fair)

[*] From now on, trainable predicated will be highlighted with an underline

[†] The framework also supports more complex Neural Network architectures, like CNN for instance.

representation of instances.

4.1.1 Knowledge base on Adult dataset

The following schema wraps up the knowledge base in the running example of a credit lending scenario, explicited on the Adult dataset from UCI-ML repository that contains 45222 instances described by 9 numerical feature with the aim of predicting whether their income exceeds $50k$\$ per year. Although it is not exactly about credit granting, we'll use the annual wage as a proxy for loan acceptance, for the sake of simplicity.

Domain

`People` (denoting the individuals from dataset)

Functions

$GoodCredit(x)$ for the real target class

To obtain information regarding the target variable, a function named *GoodCredit* is introduced, mapping each instance to its corresponding label $y \in 0, 1$. This helps in identifying two variables x_+ and x_- indicating respectively the sequence of positive and negative training examples.

Variables

x_+ for the people with good credit (positive examples).

x_- for the people with bad credit (negative examples).

x for all the people.

Fixed Predicates

$Male(x)$ privileged group made of male individuals

$Female(x)$ unprivileged group made of female individuals

Trainable Predicates

$\underline{Predict\_Granted(x)}$ for the predicted target class

---

Groundings

$\mathcal{G}(People) \subseteq \mathbb{R}^9$

$\mathcal{G}(x) = \mathbb{R}^{45.222 \times 9}$

$$\mathcal{G}(GoodCredit) : x = \begin{cases} 1, & \text{if } x \text{ was assigned the credit} \\ 0, & \text{otherwise} \end{cases}$$

$$\mathcal{G}(Male) : x = \begin{cases} 1, & \text{if } x \text{ is a male individual} \\ 0, & \text{otherwise} \end{cases}$$

$$\mathcal{G}(Female) : x = \begin{cases} 1, & \text{if } x \text{ is a female individual} \\ 0, & \text{otherwise} \end{cases}$$

$\mathcal{G}(x_+) = \langle d \in \mathcal{G}(x) \mid GoodCredit(d) = 1 \rangle$

$\mathcal{G}(x_-) = \langle d \in \mathcal{G}(x) \mid GoodCredit(d) = 0 \rangle$

$\mathcal{G}(\underline{Predict\_Granted} \mid \theta) : x \to \texttt{sigmoid}(\texttt{MLP}_\theta(x))$,

where `MLP` is a Multilayer Perceptron with learning parameters θ and single

output neuron that returns values between 0 and 1, interpreted as truth values.

Classification Axioms

$\forall x_+ \underline{Predict\_Granted(x_+)}$

$\forall x_- \neg \underline{Predict\_Granted(x_-)}$

Fairness Axioms

$\forall x(Male(x) \rightarrow \underline{Predict\_Granted(x)}) \leftrightarrow \forall x(Female(x) \rightarrow \underline{Predict\_Granted(x)})$

Learning

Learning is defined as searching the best parameters θ for the Multilayer Perceptron `MLP` that maximize the satisfability of the FOL formulas 4.5. In practice, the learning process minimizes the following loss:

$$L = (1 - _{\varphi \in \mathcal{K}} \mathcal{G}_{\theta, x \leftarrow D}(\varphi))$$

where x is grounded with the data from the dataset D.

4.2 Harnessing disparities

After expressing the classification task, we finally introduce symbolic knowledge to encode fairness constraints. As mentioned above, we focus on group fairness and specifically, we begin accounting for statistical parity. As mentioned before, this principle expresses the notion

of even distribution of resources and, in addition to convenient technical properties, it is frequently discussed in legislative contexts related to disparate impact(Feldman et al., 2015) and provides a common ground for comparison to other bias mitigation strategies. This principle is based on the assumption of independence and requires that the probability of a positive prediction, given a sensitive attribute, should be equal across all groups. Formally, this can be expressed as:

$$\mathbb{P}\{\hat{Y} = 1, A = a\} = \mathbb{P}\{\hat{Y} = 1, A = b\} \tag{4.6}$$

where $A = a, b$ corresponds to different groups identified by protected attributes whose symbolic representation needs to be encapsulated within the domain.

At this point, the knowledge base holds all the essential elements required to render the probabilistic formulation of statistical parity into a FOL axiom and we argue that it takes the form of an equivalence between two implications:

$$\forall x(Priv(x) \rightarrow \underline{C(x)}) \longleftrightarrow \forall x(Unpriv(x) \rightarrow \underline{C(x)}) \tag{4.7}$$

Intuitively, the axiom states that the truth conditions attributed to the two components of the bi-implication must have the same truth conditions and this formalization aims to ensure fairness by asserting that the predicted target $\underline{C(x)}$ is applied equally to both privileged and unprivileged groups. This addresses the research question

> RQ3 - CAN WE TRANSPOSE A STATISTICAL FAIRNESS CONSTRAINT INTO A FIRST-ORDER LOGIC AXIOM?

Although it does not seem straightforward, this representation embodies the intuition that a statistical concept can eventually be captured through a fuzzy logic axiomatization. The

Axiom	Formalization
Unconditional fair	$\forall x(Priv(x) \rightarrow \underline{C(x)}) \longleftrightarrow \forall x(Unpriv(x) \rightarrow \underline{C(x)})$
Bipartite fair	$\forall x_+(Priv(x_+) \rightarrow \underline{C(x_+)}) \longleftrightarrow \forall x_+(Unpriv(x_+) \rightarrow \underline{C(x_+)})$ $\forall x_-(Priv(x_-) \rightarrow \neg\underline{C(x_-)}) \longleftrightarrow \forall x_-(Unpriv(x_-) \rightarrow \neg\underline{C(x_-)})$
Guarded fair	$(\forall x : Priv(x))\underline{C(x)} \longleftrightarrow (\forall x : Unpriv(x))\underline{C(x)}$

Table 4.1: Different axiomatizations that were tested to optimise Demographic Parity. Only *Unconditional fair* yielded optimal results on the datasets under consideration

plausibility of this assumption can be sought in the universal quantifier: iterating over instances, its interpretation resembles a mean of clause truth value over the domain instances, thus reconciling with the probabilistic formulation in eq. 4.6. In the context of credit lending, if $\underline{Predict\_Granted(x)}$ represents the likelihood of being assigned a positive outcome and $Male(x)$ and $Female(x)$ denote two groups based on gender, the axiomatization asserts that the probability of assigning a granted label should be equivalent between these groups:

$$\forall x(Male(x) \rightarrow \underline{Predict\_Granted(x)}) \longleftrightarrow \forall x(Female(x) \rightarrow \underline{Predict\_Granted(x)})$$

By establishing this equivalence, the axiom states unbiased granted label assignment, regardless of an individual's gender.

For the sake of completeness Two other axiomatizations were devised and assessed. However, this manuscript does not contain details about formalizations that yielded worse results compared to Axiom 4.7. As illustrated in Table 4.1, these two axiomatizations share numerous similarities with the first, preserving the same logical meaning while exhibiting some distinctions:

- The *Bipartite fair axiom* enables the universal quantifier to range separately over the

positive and negative subdomains. It introduces a more fine-grained specification, giving the model an additional degree of freedom to reason about statistical parity independently within each subdomain. This approach acknowledges that privileged and non-privileged groups may be represented differently in the positive and non-positive subdomains. Preliminary results indicate that this approach does not optimise statistical parity metrics significantly.

- The *Guarded fair axiom* leverage LTN's guarded quantifier (further illustrated in section 4.2.3), where the quantifier ranges only over individuals who satisfy the internal predicate, propagating gradients exclusively across privileged and unprivileged individuals. In principle, this approach enables abstraction from the implication operator's distinct semantics, simplifying the knowledge base's definition. Preliminary results reveal that this approach optimizes fairness independence metrics, albeit with considerably worse outcomes in terms of overall fairness and trade-offs in accuracy compared to Axiom 4.7.

Bias mitigation strategy we propose in this paper strongly relies on Axiom 4.7, it is noteworthy to delve into the details concerning its implementation. We'll discuss what shall be set trainable or not, the choice of fuzzy implication operand and the interpretation of quantifiers.

4.2.1 Trainable predicates

In modeling fair classification, we leverage an *explicit and fixed grounding of constants*, which model instances with known features that must be classified, and concentrate on the key

learning task of finding the optimal grounding of logical predicates, which encode the separation of instances in groups based on their label. Observe that, since the choice of operators that interpret connectives and quantifiers deeply impacts the grounding of the formulas, this choice also impacts significantly on the satisfaction of predicates, and, eventually, on the performance of a classifier that depends on predicates' grounding.

4.2.2 Implications interpretation

Broadly speaking, implications are employed in two well-known rules of inference: *modus ponens* and *modus tollens*. Considering the implication $\forall x\, a(x) \rightarrow b(x)$, *Modus ponens* states that if $a(x)$ is known to be true, then $b(x)$ is also true. *Modus tollens* instead, poses its accent on the consequent and when $\neg\, b(x)$ is known to be true, then $\neg\, a(x)$ is true as well, this is because if the antecedent were true, the consequent should have been true as well. If in FOL the implication has a well-defined semantic, its interpretation in fuzzy logic can vary. There are two primary classes of implications generated from the fuzzy logic operators for negation, conjunction, and disjunction:

- Strong implications are defined using a fuzzy negation and fuzzy disjunction as $\alpha \rightarrow \beta = \neg\alpha \vee \beta$.

- Residuated implications are defined using a fuzzy conjunction and can be understood as a generalization of *modus ponens*, where the consequent is at least as true as the (fuzzy) conjunction of the antecedent and the implication.

When the classification predicate predicts a scenario with a false implication, there are multiple ways to rectify it. Consider the following implication from the left-hand side of Axiom

4.7:

$$\forall x(Priv(x) \rightarrow \underline{C(x)}) \tag{4.8}$$

This formula implies that all privileged examples are positive examples, namely all male individuals will receive the loan. Four categories emerge: positive privileged examples (PPE), negative non-privileged examples (NNPE), positive non-privileged examples (PNPE), and negative privileged examples (NPE). Assuming an NPE is observed, which is inconsistent with the background knowledge, there are two options:

- Modus Ponens: transfer truth from the antecedent to the consequent, namely he truth of the consequent is believed if the antecedent is true.
- Modus Tollens transfers denied truth from the consequent to the antecedent. The antecedent is believed to be false if the consequent is false.

Given that the predicate $Priv(x)$ has a fixed grounding, opting for an interpretation of implication that is modeled around *modus ponens* coluld represent a more suitable alternative. In a differentiable fuzzy logic setting, this means that when the antecedent is high, the consequent is increased. The *modus tollens* reasoning could result in being less effective since it operates by decreasing the antecedent, which cannot be changed due to its inherent fixed nature.

This might sound counter-intuitive in the first place but nevertheless ($Priv(x)$ in 4.8 represents the antecedent of the implication composing one of the sides of the double implication. Indeed, if this axiom holds true, necessarily the right consequent shall hold true as well. Assuming that the network will try to satisfy this constrains, it will modify the truth value after the training step.

Fuzzy implications, and their respective classifications, that are supported by LTN are summarised in Table 4.2. *R-Implications* involve a pure 'modus ponens reasoning' approach, while *S-Implications* incorporate both inference rules.

Implication	$I(x, y) =$	Class
Gödel I_G	$\begin{cases} 1, & \text{if } x \leq y \\ y, & \text{otherwise} \end{cases}$	R
Goguen I_P	$\begin{cases} 1, & \text{if } x \leq y \\ y/x, & \text{otherwise} \end{cases}$	R
Kleene-Dienes I_{KD}	$\max(1 - x, y)$	S
Reichenbach I_R	$1 - x + xy$	S
Łukasiewicz I_L	$\min(1 - x + y, 1)$	R + S

Table 4.2: Fuzzy implications and their classes (*R*-implications and *S*-implications).

This implication makes strong discrete choices and increases at most one of its outputs. The Gödel implication increases the consequent whenever is smaller than the antecedent, which, in turn, is never changed (Krieken et al., 2020). Moreover, as the derivative of the negated antecedent is always 0, it can never choose the *modus tollens* correction, as intended.

4.2.3 Quantifiers interpretation

Another degree of freedom lies in the aggregation function used in universal quantifiers, for which we choose the parametric A_{pME}(Badreddine et al., 2022). With this choice, universal quantifiers are represented by the generalised mean w.r.t the error, or a smooth minimum, that measures to what extent each value deviates from the ground truth. Given n truth-values

$a1, \ldots, an$ all in $[0, 1]$,

$$A_{pME}(a_1, \ldots, a_n) = 1 - \left(\frac{1}{n} \sum_{i=1}^{n} (1 - a_i)^p \right)^{\frac{1}{p}} \tag{4.9}$$

where p can be seen as a hyper-parameter as it provides flexibility in formulating more or less strict formulas, thereby accommodating or limiting the impact of outliers in data. The intuition behind the choice of p lies in that the higher its value, the higher the importance attributed to false truth values. It retrieves the arithmetic mean for $p = 1$, while the value for the quantifier starts diminishing as A_{pME} converges to the *min* operator when p increases. As already stated before, universal quantifier plays a key role in our argument since it motivates the use of fuzzy FOL to interpret a statistical quantity.

Logic Tensor Networks provide a *guarded quantifier*, an extension of the universal quantifier, to operate over a set of elements within a domain whose grounding meets a particular condition:

$$(\forall x : m(x))\varphi(x)$$

In this case, m embodies the condition, referred to as a *mask*, and $\mathcal{G}(m)$ corresponds to a function returning a Boolean value for m, meaning that 'every x satisfying $m(x)$ also satisfies $\varphi(x)$. In crisp and conventional FOL, this statement would be equivalent to $\forall x(m(x) \rightarrow \varphi(x))$, but in Real Logic, they may yield different outcomes (Badreddine et al., 2022).

4.2.4 Architecture

The architecture of our approach is summarized in Figure 4.1. Predicate $\underline{C}$ is the only that is configured to be trainable since it represents the classification task. Its parameters are trained

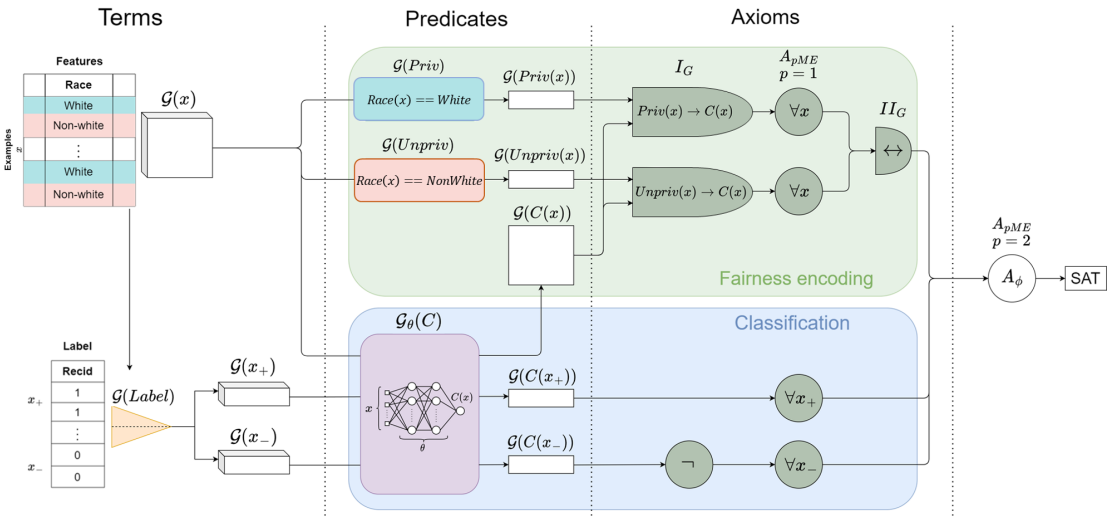


Figure 4.1: Conceptual representation of the elements of Real Logic and their role within the proposed approach. Classification task and Fairness constraints are declared separately through their respective axioms.

to jointly optimize the satisfiability of the facts in the training set (via Axiom 4.5) and the fairness constraint specified by Axiom 4.7, whose compositional interpretation is also shown in the figure.

4.3 Results on Statistical Parity

This section collects the results of the experiments and aims at evaluating to what extent our approach is able to optimise observational group fairness metrics derived from eq. 4.6 without excessively yielding on accuracy. In particular we account for Statistical Parity Difference and Disparate Impact

$$SPD = \mathbb{P}\{\hat{Y} = 1, A = a\} - \mathbb{P}\{\hat{Y} = 1, A = b\} \tag{4.10}$$

$$DI = \frac{\mathbb{P}\{\hat{Y} = 1, A = a\}}{\mathbb{P}\{\hat{Y} = 1, A = b\}} \tag{4.11}$$

for all demographic groups a, b. Optimum values for SPD are close to zero while DI implies equity for values close to one. Despite at first sight the two metrics may seem interchangeable, it is important to remark that one might be more suitable than the other depending on the specific context: in applications where the rate of positive labels is extremely low (e.g. fraud detection), minimising SPD is not an advisable choice since its value would be indeed negligible ever for very diverse values of $\mathbb{P}\{\hat{Y} = 1\}$ among groups.

Our mitigated model will be confronted with a baseline model implemented in LTN with ho fairness constraint to verify the eventual drop in accuracy while optimising non-discrimination. In addition, debiased results will be compared against SOTA in-processing approaches proposed in the literature and metrics for accuracy, SPD and DI will be checked against.

The MLP that implements the grounding of the classification predicate *underlineC* is composed of a two-layered feed-forward network with (100, 50) hidden neurons, for all the datasets. Due to their statistical nature, observational group fairness measures only make sense when calculated across samples containing a number of instances from both the sensitive groups, we approximate group fairness metrics (SPD and DI) using batch training.

All the results presented in this work have been computed through the following settings

- **Number of epochs**: 500 for all the datasets.
- **Validation**: 5-fold cross-validation.
- **Optimizer**: Adam with a fixed learning rate set to 0.001.

- **Inference**: Threshold = 0.5.
- **Classification axiomatization weight**: Fixed to 1.
- **Fairness axiom**: *Unconditional fair axiom* 4.7.
- **Quantifier aggregation**: p-mean error with a fixed $p \in \{1, 3, 5, 7\}$ for all universal aggregators in the knowledge base.
- **Fuzzy implication**: Each fuzzy operator from Table 4.2.
- **Fairness axiomatization weight**: Fixed values between 1 and 3. A value of 1 indicates that the satisfiability of the fairness axioms has the same importance as the classification axioms in the overall weighted computation of satisfiability. In contrast, a value of 3 means that the satisfiability of the fairness axioms has three times greater weight in the learning process compared to classification.

.

Our model is trained and evaluated on three benchmark datasets used in the fairness domain (Mehrabi et al., 2019), available from the UCI ML-repository:

- *Adult* income dataset has 45,222 instances. The target variable indicates whether or not income exceeds $50K per year based on census data, with gender as the protected attribute.
- *German* credit risk dataset is composed by 1,000 entries and is meant to classify bad credits based on a set of attributes encompassing demographic and financial information, including gender, that is used as a protected attribute.

	Dataset - Protected Attribute		
Metric	Adult - Gender	Compas - Race	German - Gender
SPD	0.200	0.095	0.067
DI	0.362	0.805	0.907

Table 4.3: Inherent bias over the dataset under examination

- *COMPAS* dataset includes 7,918 records collecting demographics and criminal history to predict someone's recidivism. Here, race is considered the protected attribute - restricted to white and black defendants.

The inherent bias corresponding to different metrics computed on the three datasets is reported in Table 4.3

We train our model to enforce the fairness clause formalised in Axiom 4.7, along with the classification task expressed in Axiom 4.5. The weight associated with the classifier is kept fixed, while we vary the weight of the fairness constraint: in principle, our approach allows for a fine-grained control on the degree of fairness given that the fairness axiom can be associated with a weight determining its relevance within the overall computation of the satisfiability. Hence, we investigate how accuracy and bias metrics vary according to different weights. We establish our baseline to be a plain classifier based on LTN, with no additional constraints. Being our model optimised on the probabilistic formulation of statistical parity and not on the associated metrics, we are interested in evaluating fairness both in terms of statistical parity difference and disparate impact.

In observing results, we wish to remark that enforcing fairness principles like statistical parity, which do not condition on the true target label, cannot reach perfect accuracy even with a perfect model since in that case, one would retrieve the inherent fairness level. Therefore, in

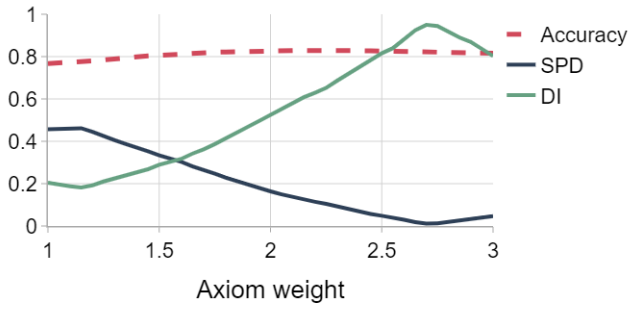


Figure 4.2: Mitigation results on Adult Income Dataset with Gödel interpretation and $p = 1$

these cases, there needs to be a trade-off between fairness and accuracy, which shall be taken into account while analysing the results from the mitigation. We perform a thorough analysis of the Adult dataset as it has a stronger inherited bias and represents the most interesting scenario to evaluate in terms of room for fairness improvements. For the remaining dataset, we report performance and fairness metrics for the optimal parameter configuration, deferring further analysis to the next Section. According to what emerges from Figure 4.2, increasing fairness axiom weight does have an impact both on SPD and DI, simultaneously, that reach almost perfect equality for weights close to 2.7. Despite the inherent bias of the Adult dataset being rather severe and one would expect a significant drop in accuracy as a result of the mitigation, the plot shows that accuracy remains relatively stable. This reveals a highly efficient optimization process: to equalise statistical parity, a significant number of predictions shall be changed, increasing the rate of positive predictions among the unprivileged group. If this relevant adjustment merely impacts accuracy, the model is mainly changing predictions to individuals that were incorrectly classified in the first place and that most likely were assigned a predicted probability close to the decision boundary.

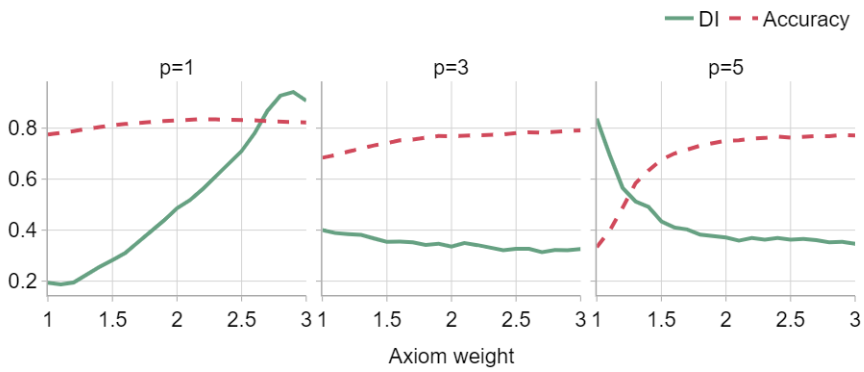


Figure 4.3: Different optimization curves for increasing value of parameter p of the universal quantifier, as a function of the fairness axiom weight. Optimal results are achieved for A_{pME} converging to arithmetic mean

Despite the architectural optimization of the network and its parameters falling outside the scope of this paper, it is noteworthy to remark that this tuning has not resulted in significant improvements in model outcomes in terms of accuracy. Indeed, the satisfiability of logical clauses - and consequently the learning process - strongly depends on the choice of the operators approximating the connectives and quantifiers. Taking into account that some first-order fuzzy semantics are better suited for gradient-descent optimization, the best-performing implementation for conjunction uses the product t-norm T_P with its dual t-conorm S_P for disjunction, together with standard negation N_s.

We also evaluate the impact of parameter p used in the A_{pME} interpretation of the universal quantifier (see eq.4.9). In a learning setting, assigning an excessively high value to p may lead to a "single-passing" operator that overly focuses on outliers at each step. This can result in gradients overfitting one input at that step, which may adversely affect the training of other inputs. This can be experimentally observed in Figure 4.3 where we observe values of DI and accuracy for different values of p. Lastly, another crucial aspect influencing the training

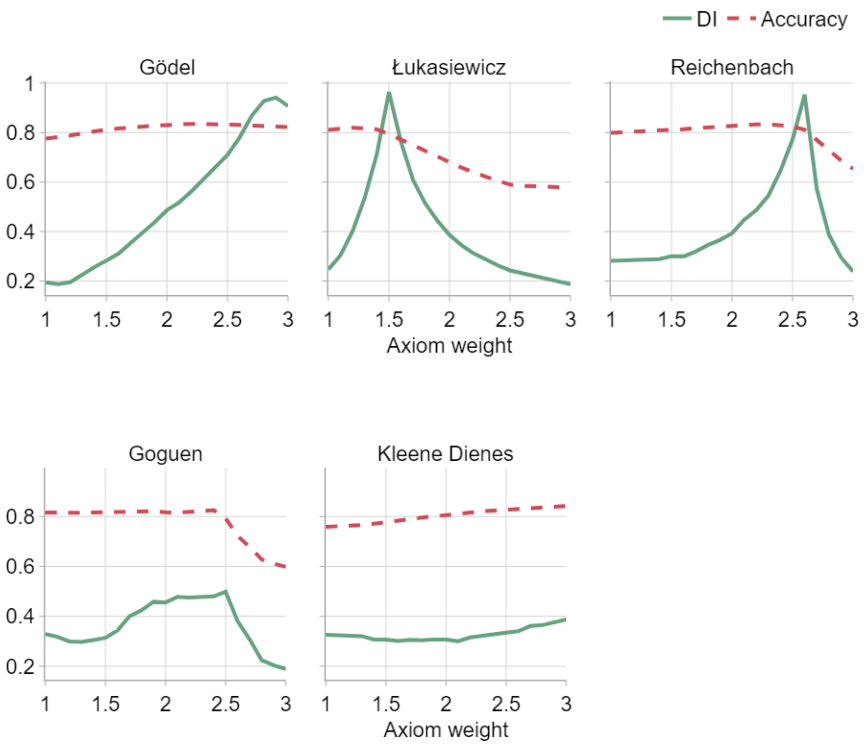


Figure 4.4: Disparate Impact and Accuracy for different implication interpretations as a function of fairness axiom weight. For both metrics, the higher corresponds to the better.

procedure involves the choice of implication operator, as thoroughly discussed above: in contrast to what reported in other contexts, we observe, that Gödel represents the optimal choice concerning the implication operator. Łukasiewicz and Reichenbach interpretations are able to reach similar Disparate Impact although slightly decreasing accuracy, while Goguen and Kleene-Dienes fail at reaching an appropriate debiasing, and will be no longer considered in further experiments.

Wrapping up, the parameters' space spans three dimensions: fairness axiom weight, implication operator and universal quantifier's exponent p. Table 4.4 reports accuracy and fairness

	Accuracy		DI		SPD	
	mitig	baseline	mitig	baseline	mitig	baseline
Adult	**0.823**	0.805	**0.949**	0.362	**0.012**	0.200
COMPAS	**0.643**	0.631	**0.957**	0.805	**0.020**	0.095
German	0.699	**0.675**	**0.966**	0.907	**0.021**	0.067

Table 4.4: Fairness metrics and accuracy on diverse dataset, choosing $p = 1$ and Gödel implication interpretation and selecting the weight that leads to the best value in term of fairness metric

metrics for all the dataset under consideration, for the optimal configuration identified by $p = 1$ and Gödel implication operator, while the optimal value for w varies according to the dataset. Results collected so far provide a positive answer to the research question

On each dataset, our approach is able to perform a mitigation that is close to complete debias, no matter the magnitude of the initial inherited bias. On mitigated predictions, the loss in accuracy is negligible, and in some cases performance increases. A possible explanation for this phenomenon is that the baseline model is keen on overfitting and the fairness enforcement act as a regulariser. It is necessary to take into account that neural-symbolic integration models offer way different advantages rather than a mere accuracy optimization.

RQ4 - CAN LOGIC TENSOR NETWORKS CORRECTLY OPTIMISE FOR FAIRNESS WITHOUT EXCESSIVELY LOOSING ON CLASSIFICATION ACCURACY?

4.3.1 Comparative results

In this section, we wish to compare our experiments with the results obtained by similar approaches that have been introduced in Section 2.3. Instead of reproducing each model, we

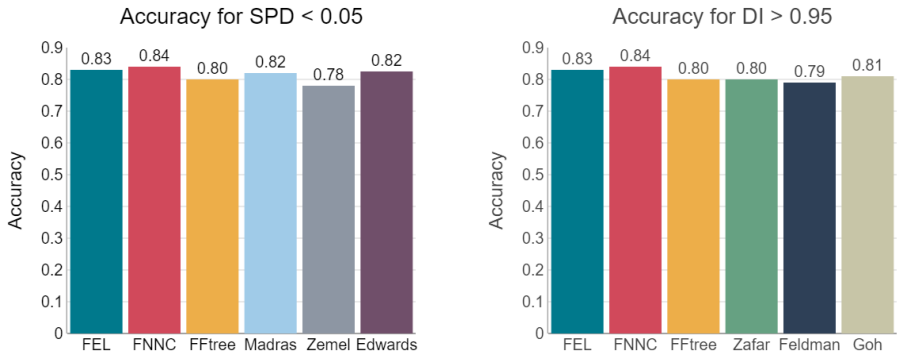


Figure 4.5: Comparative results, FEL refers to the proposed approach of Fairness Encoding in LTN. Concerning our framework, we chose the best accuracy that satisfies the imposed threshold

directly report the results from the original papers [‡]. It is noteworthy that prior studies differ from our approach in that they do not explicitly enforce SPD and DI jointly, whereas they separately formulate different optimizations. Our model, in fact, reaches the best values for the two metrics for the same input parameters and configuration, as evident from Figure 4.2. Furthermore, literature has often focused on reporting accuracy for DI > 0.8, not taking into account that inherited bias in COMPAS and German dataset already satisfy this condition, which is likely to be replicated by a non-mitigated model. Instead, in Figure 4.5 we report results at a higher threshold to better capture the behavior of our model and show it can reach way higher fairness values, much closer to perfect equality.

In comparing results, we choose the thresholds for SPD to be 0.05, higher than what reported in other works, in order to include all the approaches under consideration. This does not affect the soundness of the comparison since the chosen value is itself very close to the reachable minimum. Similarly, we deviate from the convention of evaluating accuracy for

[‡]Performances from the model from Zafar et al. (2017) and Hardt et al. (2016) are taken from the work by Donini et al. (2018) through the officially released code

DI $\geq$ 0.8 since, as discussed above, we consider it a too-mild requirement.

As evinced by Figure 4.5, when evaluated on Adult dataset, our approach of Fairness Encoding in LTN (FEL) is outperformed uniquely by FNNC (Padala & Gujar, 2020) in both metrics but is able to keep accuracy higher than any other model taken into consideration.

5

Generalising to different fairness definitions

Perhaps one of the most convenient features of applying Logic Tensor Networks to the fairness domain lies in that, at lest in principle, it can encompass diverse constraints formalisation

as it does not require a direct edit of the loss function. This section aims at answering

RQ5 - CAN OUR APPROACH BE GENERALISED TO DIFFERENT NOTIONS OF FAIRNESS?

We will be assessing whether one can declaratively account for a fairness definition that falls within the separation principle, namely in the class that equalises error rate among protected group. Indeed, this represent a test bench: conditioning on the ground truth adds new degrees of interplay among variables. The introduction of further connectives results in numerous fragments whose single satisfiability must conciliate with the more ambitious overall task.

Let us start by recalling Equal Opportunity that requires an equal predicted positive rate among groups:

$$P(\hat{Y} = 1 \mid A = a, Y = 1) = P(\hat{Y} = 1 \mid A = b, Y = 1), \qquad \forall a,\ b \in \mathcal{A}, \quad (5.1)$$

that becomes

$$\forall x((Priv(x) \wedge PositiveInstance(x)) \rightarrow \underline{C(x)}) \longleftrightarrow \forall x((Unpriv(x) \wedge PositiveInstance(x)) \rightarrow \underline{C(x)}) \quad (5.2)$$

The first-order logic formalisation highlights the utmost role of compositionality: despite a verbose appearance, eq. 5.2 only differs from 4.7 by an *and* connective on both sides that considers the conditioning on the true target label. Here, the *flow* of truth value is first computed on inner components (e.g. $Priv(x)$) and recursively calculated at outer layers, up to

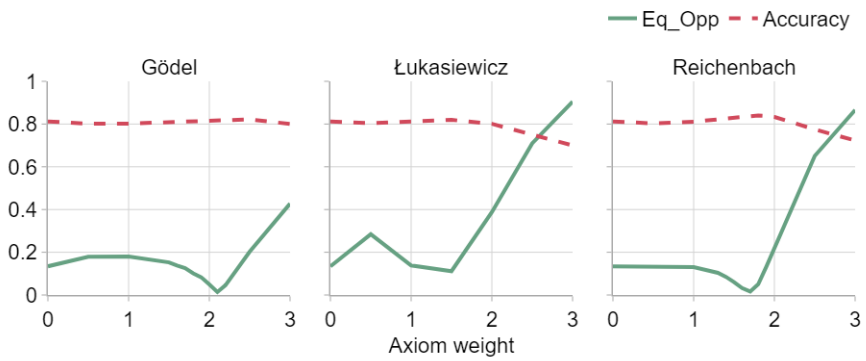
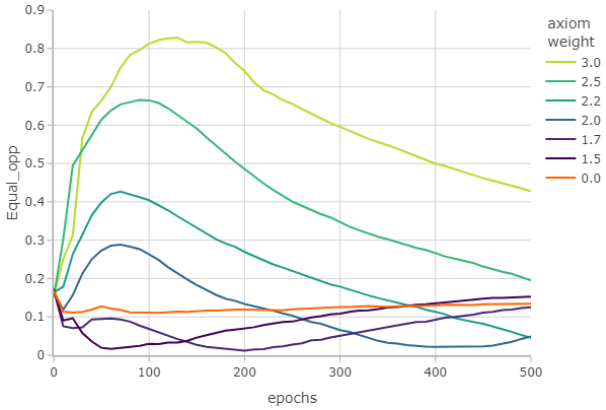


Figure 5.1: Equal Opportunity and accuracy as a function of axiom weight, each subplot refers to different implication interpretations

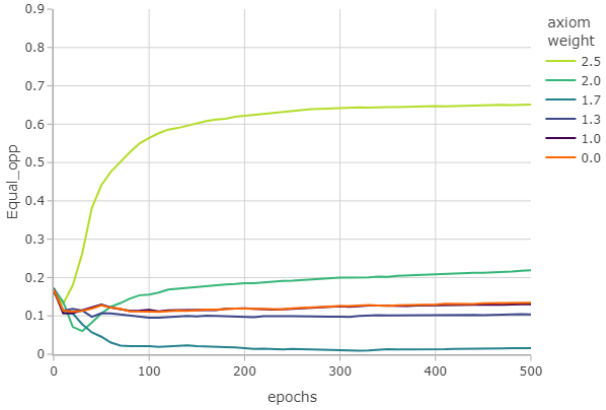
the double implication. Note that we will be measuring Equal Opportunity in terms of the difference between sides of eq. 5.1, hence the smaller, the better.

First, we run experiments on the *Adult* dataset and we measure accuracy and Equal Opportunity difference for different clause weights assigned to the fairness task. The architecture and premises are completely analogous to what illustrated in Section 4.3, and we test the three implication interpretations: Gödel, Reichenbach and Łukasiewicz. Figure 5.1 depicts that again, the model is capable of effectively carry out a debiasing – difference close to zero for Gödel and Reichenbach – while preserving the accuracy. Note that, in the context of separation principle, the situation does not necessarily demand for a trade-off between fairness and accuracy.

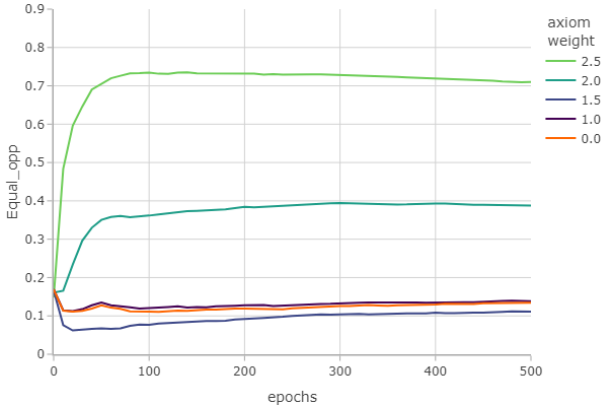
Here again, it is evident that, given a fixed number of training epochs and namely not adopting an early stopping strategy, there is a precise clause weight that represents the optimum choice for obtaining the ultimate fairness measure. However, this situation deserved a closer look and in particular Figure 5.2 depicts the behavior of the model for all implications at different training epochs for incremental weight. We immediately note that an interesting



(a) Gödel



(b) Reichenbach



(c) Łukasiewicz

Figure 5.2: Different implications on Adult dataset. The plots show how the models behave during training for increasing fairness axiom weight represented by line colors, non-mitigated models (assigned with a zero axiom weigh) highlighted in orange. Some of the weights that were tested are omitted here to avoid overlapping curves and facilitate the reading

pattern emerges in Gödel: it is not that there exist just a single optimal weight for ensuring high fairness but rather *each* weight has an optimal stopping point. Curves corresponding to incremental clause weight show an evolution that is very alike but shifted to the left as the weight increases. The same phenomenon can also be presented from a different perspective: we plot the epoch that reaches the top fairness metric as a function of the axiom weight. On the same plot but using a different scale, we report the value of predictive equality difference. Interestingly enough, for low weights, way less epochs are necessary to optimise the task, reaching the same goal while reducing the time and resource waste.

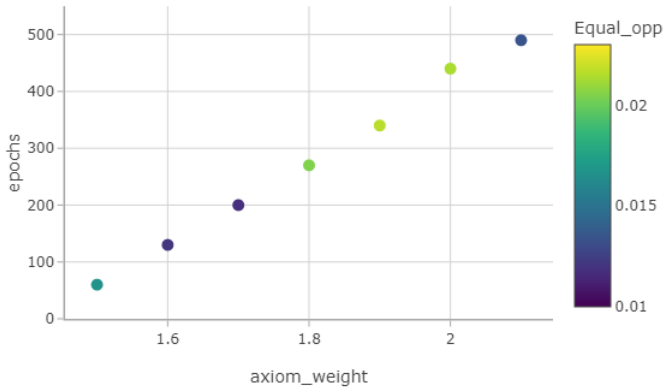


Figure 5.3: Epochs needed to perfectly optimise Equal Opportunity as a function of clause weight using Gödel implication. Marker color represent the fairness metric, although colors appear different, note that the range is extremely tiny

The remaining implications instead, do reach an asymptotic behavior and are far less dependent on the exact epoch one decide to stop the training. What the three operands have in common lies in that for large weights, the model degenerates until it fails at reaching both fairness and classification accuracy (light green lines). All in all, if one sticks to mere performance

Dataset	Implication	Accuracy		Equal Opp	
		mitig	baseline	mitig	baseline
Adult	Gödel	0.816	0.813	**0.013**	0.134
	Reichenbach	**0.840**	0.813	0.016	0.134
	Łukasiewicz$^+$	0.793	0.805	0.068	0.118
German	Gödel$^+$	0.700	0.710	0.037	0.126
	Reichenbach	0.703	0.711	**0.033**	0.172
	Łukasiewicz$^+$	**0.710**	0.713	0.063	0.165
Compas	Gödel$^+$	**0.633**	0.647	0.025	0.179
	Reichenbach$^+$	0.617	0.638	0.029	0.184
	Łukasiewicz$^+$	0.610	0.619	**0.023**	0.175

Table 5.1: Summary of results obtained on all the considered datasets and testing all implication interpretations. Experiment marked with $^+$ yielded significantly better results adopting early stopping.

at the end of training, Table A.1 wraps up quantitative results, where mitigated models are compared to baseline: they both share same architecture and parameters, and differ in that the latter only accounts for classification, namely the weight associated to the fairness axiom is set to zero. Note that, and it not a mistake, in *Adult* dataset the baseline model that only accounts for classification reaches, to the third decimal, the exact same accuracy and Equal Opportunity regardless of the implication being used.

Let us now compare our model with state-of-the-art approaches that accounts for Equal Opportunity, as wrapped up in Table 2.2 of Section 2 and we report the results in Figure 5.4. Once again, our model equals and often outperforms similar approaches presented in literature on the most common dataset used in the field and, to the best of our knowledge, no other framework reaches better results than the one we propose.

To conclude the analysis, we evaluate different implication and weight choices on the two dataset of COMPAS and *German*. On the first, the behavior at training time as well as the

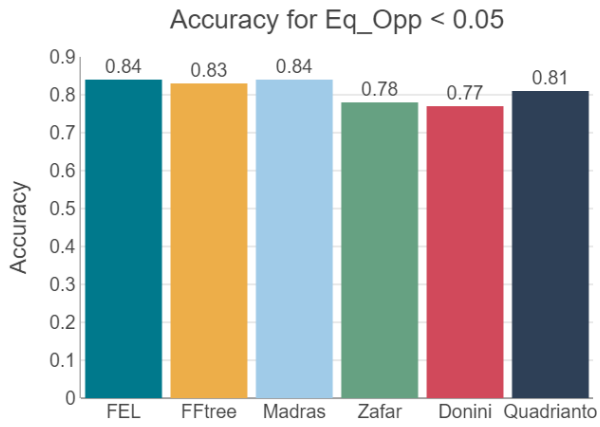


Figure 5.4: Comparative results, FEL refers to the proposed approach of Fairness Encoding in LTN.

final results mirrors what happened on *Adult*. However, on *German* dataset, despite improving on the baseline, the model exhibit a noisy behavior with large variance among different folds of the validation procedure most likely due to the much much smaller number of training (and validation) instances composing the dataset.

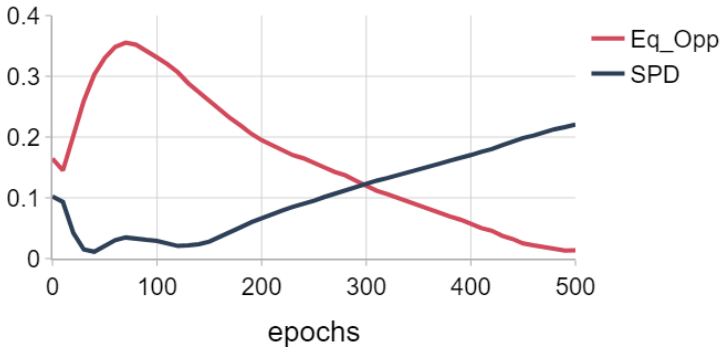


Figure 5.5: Measuring statistical parity difference while optimising for Equal opportunity, Adult dataset with Gödel implication.

The last analysis aims at checking what happens to statistical parity difference while op-

timising Equal Opportunity, to verify that the model is doing its duty on a specific metric. Results are reported in Figure 5.5 and we observe that the algorithm is indeed considering the correct constraints and discards statistical parity since it has not been instructed to take it into account in this settings.

Experimental evidence leads to the following points:

- With a light edit of the declarative axiomatization, it is possible to quick switch from a fairness notion to another
- The optimization process *does* reach parity in pretty all scenarios though in some cases early stopping significantly improved results
- Concerning the implications, Gödel is capable of optimising fairness nearly for all axiom weights, provided that one accounts for early stopping or, conversely, lets the model run for a sufficient number of epochs. Reichenbach and Łukasiewicz on the contrary, exhibit a rather stable behavior and here axiom weight plays a crucial role on the fairness metric, since there is a tiny range of weights that can optimise fairness, no matter the number of training epochs. Nonetheless, although stabilised with time, early stopping proved to be beneficial even selecting the optimal weight.
- The model considers the right fairness axiom that was declared initially, discarding other notions that were not part of this exact experiment

All in all, we can argue that Logic Tensor Networks is a poweful tool for enforcing fairness, the value of its compositionality features in enhanced by effectiveness of results and performance, paving the way to extend the application to other notions or even unseen definitions whose need might emerge from specific domain or use cases.

6

Discussion and future work

This work has collected and examined very diverse fairness notions and metrics that literature has proposed, in the attempt of disentangling their relationship and clarifying their nature. After a brief review of neuro-symbolic state-of-the art, we proposed and explored an approach that encodes fairness constraints in a binary classification setting, exploiting the declarative power of first-order logic and its fuzzy implementation in Logic Tensor Networks,

a neural-symbolic integration framework. We instill the fairness principle based on independence and, to the best of our knowledge, present the first method that remains at a higher level of abstraction and optimises on a satisfiability constraint rather than on a numerical metric. We have proposed an argument to bridge the statistical formulation of constraint with its correspondent first-order-logic axiomatization, leveraging on the fuzzy universal quantifier operator whose interpretation resembles a mean of truth value over the domain instances.

In every setting and configuration, we concurrently reach the best values for demographic parity difference and disparate impact, often at a small cost in terms of accuracy. We have also observed that in some circumstance, the biased baseline model can even be outperformed in terms of accuracy by a mitigated one, arguably because the fairness axiom acts as a regularisation and prevents overfitting on the training data. we have demonstated that adherence to non-discrimination constraint can be incrementally controlled by axiom weight, allowing to achieve the required trade-off between fairness and accuracy. We have provided a theoretical grounding of the choices in terms of universal quantifier and implication operator interpretations, which are supported by experimental evidence and provide conclusive insights on opportune choices to model fair classification with LTN, regardless of the application domain.

Experimentally, we contrast our results to similar models presented in literature, often outperforming state-of-the-art approaches, despite using a simple formalization of fairness with interpretable semantics. While we focus here on two well-known quantitative definitions of fairness, we have demonstrating to be able to effectively account for two out of the three main families of group metrics, namely independence and separation. In fact, one could consider using this framework and extending the axiomatisation to include equalised odds or predic-

tive parity for instance.

Concerning the limitations, no relevant issues have emerged while carrying out the research except perhaps the initial complexity of formalising statistical (or natural language) notions into logic axioms, which might not be the average user's cup of tea. However, once this obstacle is overcome, no further code-level adjustments to the loss function are needed. In addition, considering another generalised limitations of neuro-symbolic approaches that are considered somehow less trainable on data, we did not experience significant faults in this direction.

To conclude, further research might point the direction of exploring a different NSI framework, like LYRICS for instance, or to explore a learnable grounding for a fair representation of instances.

Acknowledgments

To start with, I need to mention the one that has literally made everything possible: my brain. It has resisted to, and sometimes actively fought, countless adversarial attacks over years and, against all odds and every my expectation, has managed to bear this tiny piece of work. So thank you brain, keep going. Continuing with irrelevant acknowledgements, here comes the greatest achievement and yet greatest failure of my entire life: our dog Bear.

My deepest gratitude goes to my amazing supervisor Prof. Matteo Palomonari for having accepted to embark this journey with me without exactly knowing what to expect. I'm glad you immediately understood what I could do and what I couldn't, and guided me towards directions where I could make an impact, even if tiny. Thank you for having provided me with just the right piece of advice in just the right moment, and I apologise for the last minute help calls, literally hours before every single deadline. Thanks to my wonderful managers Andrea and Mauro for having believed in me from the very first moment, granting me infinite degrees of freedom that made me grow and take increasing responsibilities, I'm forever

grateful. Thanks to the fantastic three of fairness: Daniele, Alessandro and Riccardo, we started knowing nothing but your enthusiasm and competence led us such far. Thanks to my colleagues and dearest friends Giulia, Stefania, Rachele and Shuyi, time spent conversing with you is never wasted.

Thanks to families, to those that provide support and to the ones who did but sadly left. From the youngest to the oldest, that cover a 90-year time span, everybody has been ready to help in all kind of situations. A special mention goes to Claudia and co. for having responded to last minute call to action several times.

And thanks to to Matteo. Life has been unkind: it fooled us showing we could reach for the stars but only provided us with a crumbling shuttle. In these years, we got trough situations that were both undoubtedly unfortunate and nerve-wrecking, I can't believe we won't be able to mend the pieces and jump-start the rocket.

Finally, I wish to thank my reviewers Prof. Salvatore Ruggieri and Prof. Luciano Serafini that promptly accepted to revise my work, your feedback has indeed improved the quality of this manuscript.

A

Additional information

This appendix contains further evidence from the experiments, in particular it shows details about the 5-fold cross-validation and takes into account the variance, in addition to the mean.

Dataset	Implication	Accuracy		Equal Opp	
		mitig	baseline	mitig	baseline
Adult	Gödel	0.816 ± 0.004	0.813 ± 0.003	$\mathbf{0.013 \pm 0.011}$	0.134 ± 0.028
	Reichenbach	$\mathbf{0.840 \pm 0.002}$	0.813 ± 0.003	0.016 ± 0.012	0.134 ± 0.028
	Łukasiewicz$^+$	0.793 ± 0.005	0.805 ± 0.003	0.068 ± 0.020	0.118 ± 0.022
German	Gödel$^+$	0.700 ± 0.022	0.710 ± 0.031	0.037 ± 0.023	0.126 ± 0.043
	Reichenbach	0.703 ± 0.026	0.711 ± 0.034	$\mathbf{0.033 \pm 0.028}$	0.172 ± 0.119
	Łukasiewicz$^+$	$\mathbf{0.710 \pm 0.030}$	0.713 ± 0.035	0.063 ± 0.025	0.165 ± 0.065
Compas	Gödel$^+$	$\mathbf{0.633 \pm 0.006}$	0.647 ± 0.007	0.025 ± 0.025	0.179 ± 0.022
	Reichenbach$^+$	0.617 ± 0.007	0.638 ± 0.006	0.029 ± 0.023	0.184 ± 0.027
	Łukasiewicz$^+$	0.610 ± 0.007	0.619 ± 0.008	$\mathbf{0.023 \pm 0.012}$	0.175 ± 0.017

Table A.1: Summary of results for Equality of Opportunity obtained on all the considered datasets and testing all implication interpretations with evidence of the standard deviation. Experiment marked with $^+$ yielded significantly better results adopting early stopping.

References

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

Bader, S. & Hitzler, P. (2005). Dimensions of neural-symbolic integration - A structured survey. *CoRR*, abs/cs/0511042.

Bader, S., Hitzler, P., & Hölldobler, S. (2004). The integration of connectionism and first-order knowledge representation and reasoning as a challenge for artificial intelligence. *arXiv preprint cs/0408069*.

Badreddine, S., d'Avila Garcez, A., Serafini, L., & Spranger, M. (2022). Logic tensor networks. *Artificial Intelligence*, 303, 103649.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Barocas, S. & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104, 671.

Berk, R. (2009). The role of race in forecasts of violent crime. *Race and social problems*, 1(4), 231.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth, A. (2017). A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

Binns, R. (2020). On the apparent conflict between individual and group fairness. (pp. 514–524).

Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91).: PMLR.

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building classifiers with independency constraints. *ICDM Workshops 2009 - IEEE International Conference on Data Mining*, (pp. 13–18).

Castelnovo, A., Cosentini, A., Malandri, L., Mercorio, F., & Mezzanzanica, M. (2022a). Fftree: A flexible tree to handle multiple fairness criteria. *Information Processing & Management*, 59, 103099.

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022b). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5.

Cleary, T. A. (1966). Test bias: Validity of the scholastic aptitude test for negro and white students in integrated colleges. *ETS Research Bulletin Series*, 1966.

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, 5.

Cole, N. S. (1973). Bias in selection. *Journal of educational measurement*, 10(4), 237–255.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Part F129685.

Crenshaw, K. W. (1994). Mapping the margins. *The public nature of private violence*, (pp. 93–118).

Darlington, R. B. (1971). Another look at "cultural fairness" 1. *Journal of educational measurement*, 8(2), 71–82.

D'Avila Garcez, A., Besold, T. R., De Raedt, L., Földiak, P., Hitzler, P., Icard, T., Kühnberger, K.-U., Lamb, L. C., Miikkulainen, R., & Silver, D. L. (2015). Neural-symbolic learning and reasoning: contributions and challenges. *2015 AAAI Spring Symposium Series*.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., & Pontil, M. (2018). Empirical risk minimization under fairness constraints. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 31.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *ITCS 2012 - Innovations in Theoretical Computer Science Conference*.

Edwards, H. & Storkey, A. (2015). Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015-August.

Fleisher, W. (2021). What's fair about individual fairness? *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. *Association for Computing Machinery*.

Garcez, A. D. & Lamb, L. C. (2020). *Neurosymbolic AI: The 3rd Wave*.

Goh, G., Cotter, A., Gupta, M., & Friedlander, M. P. (2016). Satisfying real-world goals with dataset constraints. *Advances in neural information processing systems*, 29.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, (pp. 3323–3331).

Hertweck, C., Heitz, C., & Loi, M. (2021). On the moral justification of statistical parity. (pp. 747–757).

Hitzler, P., Eberhart, A., Ebrahimi, M., Sarker, M. K., & Zhou, L. (2022). Neuro-symbolic approaches in artificial intelligence. *National Science Review*, 9(6), nwac035.

Hutchinson, B. & Mitchell, M. (2019). 50 years of test (un) fairness: Lessons for machine learning. (pp. 49–58).

Joseph, M., Kearns, M., Morgenstern, J. H., & Roth, A. (2016). Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29.

Kahneman, D. (2017). *Thinking, fast and slow*. New York :Farrar, Straus and Giroux.

Kamiran, F. & Calders, T. (2009). Classifying without discriminating. *2009 2nd International Conference on Computer, Control and Communication, IC4 2009*.

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. *Proceedings - IEEE International Conference on Data Mining, ICDM*.

Kamiran, F., Žliobaitė, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems*, 35(3), 613–644.

Kamishima, T., Akaho, S., & Sakuma, J. (2011). Fairness-aware learning through regularization approach. *Proceedings - IEEE International Conference on Data Mining, ICDM*, (pp. 643–650).

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning* (pp. 2564–2572).: PMLR.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2019). An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 100–109).

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *Association for Computing Machinery*.

Kolata, G. (1982). How can computers get common sense? two of the founders of the field of artificial intelligence disagree on how to make a thinking machine. *Science*, 217(4566), 1237–1238.

Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., & Kompatsiaris, Y. (2018). Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*.

Krieken, E. V., Acar, E., & Harmelen, F. V. (2020). Analyzing differentiable fuzzy implications. *17th International Conference on Principles of Knowledge Representation and Reasoning, KR 2020*, 2.

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research*, 43(2), 139–161.

Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., & Puri, R. (2019). Bias mitigation post-processing for individual and group fairness. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2019-May.

Louizos, C., Swersky, K., Li, Y., Welling, M., & Zemel, R. (2016). The variational fair autoencoder. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*.

Luong, B. T., Ruggieri, S., & Turini, F. (2011). k-nn as an implementation of situation testing for discrimination discovery and prevention. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2018). Learning adversarially fair and transferable representations. *35th International Conference on Machine Learning, ICML 2018*, 8.

Marcus, G. (2020). The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.

McClelland, J. L. & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature reviews neuroscience*, 4(4), 310–322.

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5, 115–133.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *Association for Computing Machinery*.

Menon, A. K. & Williamson, R. C. (2018). The cost of fairness in binary classification. *Proceedings of Machine Learning Research*, 81, 1–12.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8, 141–163.

Padala, M. & Gujar, S. (2020). Fnnc: Achieving fairness through neural networks. *IJCAI International Joint Conference on Artificial Intelligence*, 2021-January.

Petersen, N. S. & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, (pp. 3–29).

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in Neural Information Processing Systems*, 2017-December.

Quadrianto, N. & Sharmanska, V. (2017). Recycling privileged learning and distribution matching for fairness. *Advances in neural information processing systems*, 30.

Räz, T. (2021). Group fairness: Independence revisited. (pp. 129–137).

Rogers, T. T. & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

Roy, A. (2011). A theory of the brain - the brain uses both distributed and localist (symbolic) representation. *The 2011 International Joint Conference on Neural Networks*, (pp. 215–221).

Serafini, L. & d'Avila Garcez, A. S. (2016). Learning and reasoning with logic tensor networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10037 LNAI.

Simoiu, C., Corbett-Davies, S., & Goel, S. (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3), 1193–1216.

Sun, Y., Fung, B. C., & Haghighat, F. (2022). In-processing fairness improvement methods for regression data-driven building models: Achieving uniform energy prediction. *Energy and Buildings*, 277.

Thanh, B., Ruggieri, S., & Turini, F. (2011). : (pp. 502–510).

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement*, 8(2), 63–70.

Verma, S. & Rubin, J. (2018). Fairness definitions explained. *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, (pp. 1–7).

Wagner, B. & d'Avila Garcez, A. (2021). Neural-symbolic integration for fairness in ai. *CEUR Workshop Proceedings*, 2846.

Wang, W. & Yang, Y. (2022). Towards data-and knowledge-driven artificial intelligence: A survey on neuro-symbolic computing. *arXiv preprint arXiv:2210.15889*.

Washington, A. L. (2018). How to argue with an algorithm: Lessons from the compas-propublica debate. *Colo. Tech. LJ*, 17, 131.

Yang, F., Cisse, M., & Koyejo, O. O. (2020). Fairness with overlapping groups; a probabilistic perspective. *Advances in Neural Information Processing Systems*, 33.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. (pp. 1171–1180).

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2019). Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20, 1–42.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *30th International Conference on Machine Learning, ICML 2013*.

Zhang, Z., Wang, S., & Meng, G. (2023). *A Review on Pre-processing Methods for Fairness in Machine Learning*. Association for Computing Machinery.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 12 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The original template was released under the permissive MIT (X11) license from its author Jordan Suchow. .

Experiments were conducted using an Intel Core i7 equipped with GPU Nvidia Quadro, the code was written in python 3.8 exploiting the Tensorflow official image deployed into a Docker container. Results were collected trough Weight and Bias platform and plots were made using Plotly.