Support Fair Decisions by Actionable Ethics through Neural Learning: an Evaluation over COMPAS

Luca Squadrone, Danilo Croce and Roberto Basili

Department of Enterprise Engineering
University of Rome, "Tor Vergata"

{croce,basili}@info.uniroma2.it

Abstract

AI technology is increasingly adopted in different fields, such as in criminal justice. This is emphasizing the role of algorithms supporting decision-making processes able to limit the impact of ethical bias. Focusing on COMPAS, a largely adopted resource in ethical AI studies, we present experiments of the so-called *ethics-by-design* paradigm. We suggest that integrating ethical principles in the learning process of neural architecture to optimize the balance between business accuracy and ethics is viable. We assess the resulting fairness through the balance of False positive rates (*Predictive equality*) and False negative rates (*Equal opportunity*).

1 Introduction

Machine learning applications exhibit exponential growth, now implemented in high ethical risk scenarios, such as granting a loan, hiring decisions, prediction of criminal recidivism [Savani *et al.*, 2020]. Clear advantages of using machine learning algorithms include the ability to quickly and accurately analyze large amounts of data. However, this opens the way to algorithms generating discriminatory predictions against individuals or social groups [Buolamwini e Gebru, 2018; Snow, 2018; Édouard Richard e Schacht, 2020; Ali *et al.*, 2019; Angwin e Kirchner, 2016; Osoba e Welser IV, 2017], as for the prejudice inherent in the ways historical data have been collected. This phenomenon, known as *bias*, makes data management, able to rebalance ethically distorting phenomena, a critically important technology.

Let us consider, for example, COMPAS, a system used as a support tool by judges to predict the risk of a defendant's recidivism. It has been found that black defendants were predicted with a higher risk of recidivism than their actual risk compared to white defendants due to historical data.

Since the algorithms are based on mathematical and statistical principles, they are unable to independently recognize the ethical values related to skin color or gender difference. Consequently, it is necessary to introduce a correcting force. This is further highlighted by recent studies [Buolamwini e Gebru, 2018; Snow, 2018; Édouard Richard e Schacht, 2020; Ali *et al.*, 2019; Angwin e Kirchner, 2016; Lambrecht e Tucker, 2019; Ingold e Soper, 2016] which have shown how machine learning algorithms may emphasize human factors such as prejudices, cliches and errors of assessment. Therefore, it becomes of fundamental importance that machines are somehow able to take decisions aligned with human values and expectations to avoid the risk of dangerous drifts in terms of ethics and human values.

In this paper, we have focused on how to mitigate this type of bias. Basically, we can say that an artificial intelligence system is really useful only if it can make decisions that respect ethical principles, especially when dealing with "human" issues as delicate as (apparently) alien to the mere algorithmics. Much of the discussion around discrimination and inequality risks in ML focuses on various solutions to reduce "bias" in algorithms, such as modifying training data or diversifying training data sources to reduce disparities between groups ([Feldman *et al.*, 2015; Hardt *et al.*, 2016; Iosifidis *et al.*, 2019]). However, research such as [Lungren, 2021] suggests that such approaches may fail when it is difficult isolating protected attributes from data. In particular, we introduce the idea of in-processing ethics by design [1] which, unlike the previous ones, focuses on the learning stage. In this way, it makes ethical learning possible even in situations where the ethical correlation between input-output is unknown. As extensively discussed in [Caton e Haas, 2020], [Pessach e Shmueli, 2020] and [Mehrabi *et al.*, 2021], in processing methods are generally used to mitigate biases in the algorithms. These methods fall under three categories: *pre-processing*, *in-processing*, and *post-processing* algorithms. Pre-processing methods manipulate the training dataset *before* training a model, in-processing methods modify the learning machine itself while post-processing techniques modify decisions made by a given machine. Our methodology corresponds to an in-processing method as we discuss in the following section.

2 Methodology

The core of our approach ([Rossini *et al.*, 2020]) is to model ethics via automated reasoning over formal descriptions, e.g. ontologies, but by making it available *during the learning stage*. In fact, we give for granted an explicit formulation of

[1] Ethics by Design (EbD) is an approach to design that aims at the systematic inclusion of ethical values, principles, requirements and procedures since the design and training stages of an ML process.

ethical principles and focus on a form of ethical learning, as *external alignment* (learning from others, [Kleiman-Weiner *et al.*, 2017]). Here, ethical evidence is inferred from an ethical ontology and used to guide the model selection of a deep learning setting. The resulting deep neural network jointly models the functional as well as the ethical conditions characterizing the underlying decision making. In this way, the discovery of latent ethical knowledge, i.e., hidden information in the data that is meaningful under the ethical perspective, is enabled and made available to the learning process. The target of this framework is a learning machine able to select the best decisions among those that are also ethically sustainable.

Ethical principles. The abstract ethical principles are enforced through *Ethical Rules* that constraint individual features (e.g. gender) and determine the degree of ethicality of principles over their domains. Ethical Rules usually target (i.e. define and manipulate) one or more features and assign values (or better, establish some probability distributions) across the feature domains. These rules are termed as *truth-makers* ($\mathcal{TM}$) as they account the possibly uncertain ethical state of the world regarding decisions over the individual instances.

The role of truth-makers. Truth-makers are rules of the $\mathcal{EO}$ ontology that actively determine the ethical profile of a decision $d(i)$ over an instance i. In particular, given a pair $\big(i, d(i)\big)$, a truth-maker tm will determine a probability distribution to the set of ethical benefit and ethical risk dimensions. For every tm, ethical dimension $e_j(i)$ and possible ethical value $v_k \in V$ the following probability is defined:

$$P\Big(e_j(i) = v_k \mid \big(\vec{i}, d(i)\big), tm\Big) \qquad \forall j, \forall k = 1, \ldots, 5$$

which expresses the evaluation of the truth-maker tm onto the representation $\vec{i}$ of an instance i, given the decision $d(i)$ and along the k-th value of the j-th ethical dimensions. A truth-maker thus assigns probabilities to the ethical signature of an individual i for all possible combinations of business characteristics $\vec{f}(i)$ and decisions $d(i)$[2]. Multiple truth-makers can contribute to a given ethical feature $e_j(i)$ individually biasing their overall probability $P(e_j(i))$. When all truth-makers are fired, the resulting *ethical signature* over an instance $\vec{i}$ and its decision $d(i)$ consists $\forall j, k$:

$$es_j(i) = \prod_{tm} P\Big(tm|\mathcal{EO}\Big) P\Big(e_j(i) = v_k \mid \big(\vec{i}, d(i)\big), tm\Big)$$

Architecture. The network consists of several components, trained in 3 different phases and each with a specific loss function (Figure 1). In the *first phase*, an ethical encoding network is defined, responsible for learning the combinations of input features that capture possible relationships between business observations and undesired ethical consequences: these latter are coded as ethical features. In a *second* stage, Business Expert and Ethics Expert are activated, whose role is respectively to estimate *independently* the distributions of suitable business decisions, on the one side, and predict as well their ethical consequences. In a *third* stage, an Ethical-aware Network is responsible for estimating the joint probability of each possible triplet (decision, benefit, risk), which associates the risks and ethical benefits triggered by a certain decision to each instance. This third phase concludes with the *final business decision* of the network that is determined by a decision policy over risks and opportunities.

Once a new instance is available the decision is derived as a combination of business policies, implemented by a network as the decision that maximizes analogy with past (i.e. training) cases, and ethical control, aimed at maximizing benefits and minimizing risks. We model both policies as different loss functions used to train specialized sub-networks, called business and ethical expert. Most interestingly, this architecture formulation allows emphasizing the contribution of each triple in the probability estimation through a factor β (the exponential Tweaking factor in [Rossini *et al.*, 2020]), so that we can train the overall network by balancing business accuracy and ethical compliance. Emphasis on ethical consequences can be achieved by amplifying ethics constraints, i.e. by tweaking β to larger values.

While gold standards usually provide just business decisions, that are defined as known, i.e. sharp probability distributions across possible outcomes, these are not guaranteed to be ethical. Smoothing is needed to allow the neglecting of unethical cases so that decisions d_i different from the gold standard decisions are given non-zero probabilities. *Laplace smoothing* is applied across the K different classes, expressed by: $\hat{d}_i = \frac{d_i + \alpha}{1 + K\alpha}$.

Two policies have been defined: (1) the final decision is derived only from the probability triplets that respect the ethical constraints; (2) the second, on the other hand, does not impose ethical constraints, and is derived simply by summing up all probability contributions of all triplets.

For more details, refer to the paper [Rossini *et al.*, 2020]

3 Experimental analysis

Now let's see how, using the ethics-by-design method just described, it is possible to achieve the goal of reducing the problems related to bias.

One of the most famous algorithmic bias scenarios in the world is COMPAS, on which ethical problems regarding the discrimination of the African-American population have been identified. COMPAS[3] is a software owned by Northpointe (now Equivant[4]). It is a support tool used in criminal justice to predict the risk of a defendant's recidivism. The aim is to support the judge in deciding whether or not to release defendants awaiting trial on bail, where they are deemed potentially too dangerous. In 2015, the independent American organization ProPublica demonstrated how this software makes its decisions on a discriminatory basis, in particular towards the African-American population [Angwin e Kirchner, 2016]. In

[2] If no truth-maker is triggered by an instance the uniform probability distribution u is used, i.e. $P\Big(e_j(i) = v_k|\big(\vec{i}, d(i)\big), tm\Big) = \frac{1}{m}$, over the values v_k and different ethical features, i.e. $\forall j, k$.

[3] Acronym for Correctional Offender Management Profiling for Alternative Sanctions

[4] https://www.equivant.com

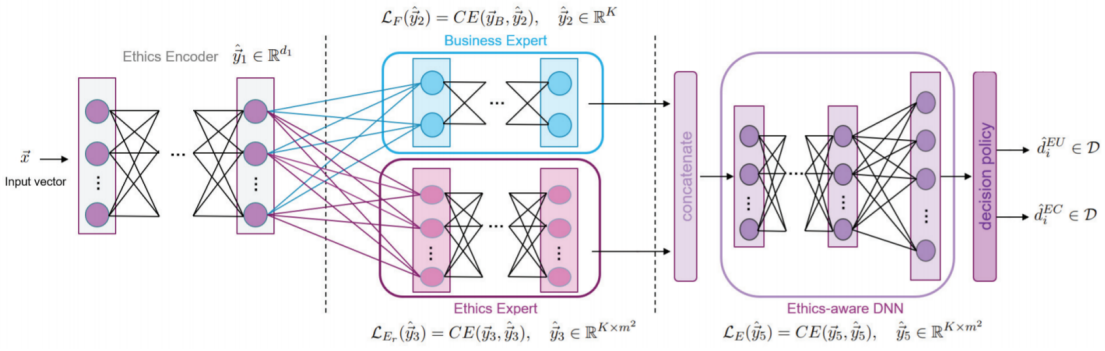

Figure 1: Network Architecture Proposed in (Rossini *et al.*, 2020).

β	*Acc*	*EthCompl*	*non Afr.-Amer.* FPR	*non Afr.-Amer.* FNR	*Afr.-Amer.* FPR	*Afr.-Amer.* FNR
- *(MLP)*	0.72	0.69	0.19	0.48	0.39	0.25
0.01	0.70	0.69	0.15	0.53	0.38	0.23
0.10	0.68	0.76	0.10	0.63	0.27	0.37
0.11	0.68	0.77	0.09	0.65	0.24	0.39
0.12	0.67	0.78	0.09	0.66	0.23	0.41
0.13	0.67	0.78	0.09	0.67	0.23	0.43
0.14	0.66	0.80	0.09	0.66	0.20	0.45
0.15	0.66	0.81	0.08	0.69	0.19	0.48

Table 1: Results by varying the parameter β on the COMPAS dataset. Values express the average over 5 different runs.

our opinion, the approach we designed represents an effective improvement over the resolution of the observed discrimination. Our analysis is carried out over the dataset published by ProPublica[5] extracted from the data used by COMPAS.

Given the architecture discussed above, we applied some ethical principles to the problems outlined by literature about COMPAS, namely the bias related to race and sex. Given the ethical principles, implemented as feature-specific truth-makers, we estimated the ability of the system to increase ethical compliance and to impact the results through a significant reduction in bias. In order to evaluate the ethical system under different working conditions, various smoothing and tweaking factor settings have been applied $(\alpha, \beta) \in \{0.1,\ 0.3,\ 0.5\} \times \{0.01,\ 0.1,\ 0.11,\ 0.12,\ 0.13,\ 0.14,\ 0.15\}$ to systematically study the dependence of the benefit from the greediness of the system w.r.t. to the sensitivity to ethical principles. During the validation phase, for each tweaking value, our system compares all the models as the smoothing factor changes, in order to identify the best among them. Remember that higher values for β correspond to more influential ethical losses during training. More in detail, for each β we choose the smoothing value α maximizing the accuracy on a validation set. To assess whether or not the decisions made by our model are also respecting the ethical ontology in use, a second measure, namely *Ethical Compliance* ($EthCompl$), is computed as $\frac{\mathcal{D}^+}{\mathcal{D}^+ + \mathcal{D}^-}$, where D^+ represents the number of ethically compliant instances and D^- the non-compliant ones.

We modeled ethical principles on this dataset through three different truth-makers, that act upon the attributes of the dataset sensitive to discrimination (i.e. $race$, sex and age). The truth maker are designed to favor a fair balancing of the ratio between False Positive rates (FPR[6]) and False Negative rates (FNR[7]). To give an idea about the working of truth makers, we hereafter report the definition for the race feature, only. More details are in [Croce *et al.*, 2022].

The $race$ truth-maker operates on two factors that attempt to increase risks and, at the same time, to reduce benefits on the assessments of recidivism regarding African Americans. The truth maker should not act on other races and leave all decisions not involving African Americans unaffected. We can describe these in high-level language as follows:

1. Increase the benefit of classifying an African American as a non-repeat offender.

   Decrease the risk of classifying an African American as a non-repeat offender.

   Decrease the benefit of classifying an African American as a repeat offender.

   Increase the risk of classifying an African American as a repeat offender.

2. Do not act on anyone who is not African-American.

In Table 1, the first line reports the performance of a system trained through a simple multilayer perceptron, that has no sensitivity to the ethical dimension of the problem. Its accuracy 0.72 is counterbalanced by poor ethical compliance (only 69% of the instances can be considered ethically acceptable given our principles). As we can see in the rest of the Table 1, increasing the parameter β, that is the influence of the ethical component in the network, the $EthCompl$ increases. At the same time, the inequality of treatment towards African-Americans also decreases as shown by the comparable false positive rates of the two protected groups[8]. Although

[5] We used "compas-scores-two-years.csv" published on https://github.com/propublica/compas-analysis. The dataset is made up of 7,214 instances

[6] $FPR = \frac{FP}{\text{Actual Negative}} = \frac{FP}{TN+FP}$

[7] $FNR = \frac{FN}{\text{Actual Positive}} = \frac{FN}{TP+FN}$

[8] African-Americans concerning other races

this inevitably involves a small loss of accuracy, it compensates by a largely more ethical decision (0.81 vs. 0.69).

Notice that tweaking allows to identify the optimal balance as an operationally cost-effective compromise between the business and ethical performance of the system. The goal was achieved thanks to cross-validation aimed at studying the behavior of the *Accuracy* and *EthCompl* scores according to varying values for β.

4 Conclusions

Experiments around the ethics by design framework discussed by [Rossini *et al.*, 2020] confirm its ability to strongly reduce the imbalance in a strongly biased data set such as COMPAS. The results are much better judgment over African Americans with basically no change on anyone else. This result was achieved at the expense of a more than acceptable loss of performance, which in our opinion is a very significant result.

Possible future extensions of the work carried out are possible, such as the identification of the strategies adopted by the model to produce a certain better decision. This result could be obtained by identifying which features were initially used and how their importance changes when the ethical relevance is increased. Along this line, tools for model explainability, such as LIME [Ribeiro *et al.*, 2016], can be used as a tool for tracking features involved in the ethical ontology change importance when β changes. More specifically, we can expect to guarantee the observation of protected variables that may lose relevance during the ethical upgrading of the model decision process. We also think that this process can be used to track potential dependencies between protected variables (race, sex, age) and other related variables (for example, the postal code).

References

[Ali *et al.*, 2019] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, e Aaron Rieke. Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–30, 2019.

[Angwin e Kirchner, 2016] Mattu Angwin, Larson e ProPublica Kirchner. Machine bias, 2016.

[Buolamwini e Gebru, 2018] Joy Buolamwini e Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[Caton e Haas, 2020] Simon Caton e Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

[Croce *et al.*, 2022] Danilo Croce, Luca Squadrone, e Roberto Basili. Evaluating actionable ethics on the smoothing of biased phenomena for fair machine learning. In *Forthcoming*, 2022.

[Feldman *et al.*, 2015] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, e Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the KDD 2015*, pages 259–268, 2015.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, e Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

[Ingold e Soper, 2016] David Ingold e Spencer Soper. Amazon doesn't consider the race of its customers. should it?, 2016.

[Iosifidis *et al.*, 2019] Vasileios Iosifidis, Besnik Fetahu, e Eirini Ntoutsi. Fae: A fairness-aware ensemble framework. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1375–1380. IEEE, 2019.

[Kleiman-Weiner *et al.*, 2017] Max Kleiman-Weiner, Rebecca Saxe, e Joshua B. Tenenbaum. Learning a commonsense moral theory. *Cognition*, 167:107–123, 2017.

[Lambrecht e Tucker, 2019] Anja Lambrecht e Catherine Tucker. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management science*, 65(7):2966–2981, 2019.

[Lungren, 2021] Matthew Lungren. Risks of ai race detection in the medical system, 2021.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, e Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[Osoba e Welser IV, 2017] Osonde A Osoba e William Welser IV. *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation, 2017.

[Pessach e Shmueli, 2020] Dana Pessach e Erez Shmueli. Algorithmic fairness. *arXiv preprint arXiv:2001.09784*, 2020.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, e Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the KDD 2016*, pages 1135–1144, 2016.

[Rossini *et al.*, 2020] Daniele Rossini, Danilo Croce, Sara Mancini, Massimo Pellegrino, e Roberto Basili. Actionable ethics through neural learning. In *Proceedings of the AAAI 20202*, pages 5537–5544, 2020.

[Savani *et al.*, 2020] Yash Savani, Colin White, e Naveen Sundar Govindarajulu. Intra-processing methods for debiasing neural networks. *arXiv preprint arXiv:2006.08564*, 2020.

[Snow, 2018] Jacob Snow. Amazon's face recognition falsely matched 28 members of congress with mugshots, 2018.

[Édouard Richard e Schacht, 2020] Nicolas Kayser-Bril Édouard Richard, Judith Duportail e Kira Schacht. Undress or fail: Instagram's algorithm strong-arms users into showing skin, 2020.