Addressing algorithmic fairness through metrics and explanations

IDAI 2021 Summer School
Course 5 (11:30-13:00)

Miguel Couceiro

Université de Lorraine, CNRS, Inria Nancy-Grand Est, LORIA

Outline

- Tackling data biases: balancing datasets
- Using explanations for assessing process fairness
- Addressing unfairness through unawareness
- Some use cases and available resources
- Discussions

Tackling data biases

Biased Data

Unbalanced dataset

- Common in real-world datasets
- Categories are not equally represented
- Lead to misclassification of under-represented categories

SMOTE

- **S**ynthetic **M**inority **O**versampling **TE**chnique[1]
- Over-sampling minority class by creating "synthetic" examples

[1] Chawla, *et al.* Synthetic Minority Over-sampling Technique. JAIR. 2002

Using explanations for assessing fairness

Recall...

Based on **decision outcomes**, fairness can be assessed based on:

- **Fairness metrics**: individual & group fairness, equal opportunity, demographic parity, equal accuracy, etc.
- **Process fairness**: model's dependence on "sensitive features" (e.g., salient features such as race, age, or sex,...)

Two main approaches to dealing with ML unfairness:

1. Enforce fairness constraints while learning, e.g.:

   $$P(y_{\text{pred}} \neq y_{\text{true}} | race = Black) = P(y_{\text{pred}} \neq y_{\text{true}} | race = White)$$

   Drawback: Complexity, fairness "gerrymandering" & overfitting

2. Exclude sensitive/salient features (for instance, COMPAS)

   Drawback: Decreased accuracy!

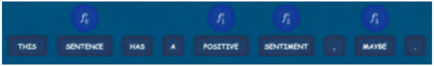
Idea: Use FI-explanations to measure dependence on "sensitive features"

Local explainers: Simple surrogates on a neighbourhood

These frameworks are based on three main components:

- Interpretable Data Representation: two-way translation $x \mapsto z_x$ of the orginal data into (and from) an interpretable domain.
- Data Sampling: choice of neighboorhood of the instance to explain
- Explanation Generation: learning the surrogate (often linear) on the chosen neighbourhood in the interpretable domain. Weights give FI.

$z_x = [0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 0]$



$z_x = [1, 1, 1, 1]$

https://github.com/fat-forensics/events/blob/master/resources/2020_ecml-pkdd/slides/1.2-surrogates.pdf

LIME: Local Interpretable Model-agnostic Explanations[2]

LIME: learns a linear $g \in \mathcal{G}$ on a neighborhood of z_x (x to explain) by

$$g = argmin_{g' \in \mathcal{G}} \mathcal{L}(f, g', \pi_{z_x}) + \Omega(g')$$

for the distance $\mathcal{L}(f, g', \pi_{z_x})$ of f and g' on the kernel π_{z_x}

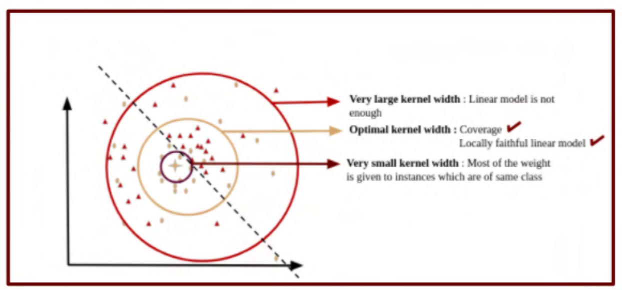


Figure 1: Illustration of optimal kernel on the (interpretable) space (z_x's)

[2] Ribeiro, *et al.* "Why Should I Trust You?": Explaining predictions of any...

LIME Explanations[3]

LIME: learns a model g on the neighborhood of z_x to explain

$$g(z_x) = \alpha_0 + \sum_{1 \leq i \leq d'} \alpha_i z_{x_i},$$

where $\hat{\alpha}_i$ represents the contribution or importance of feature z_x

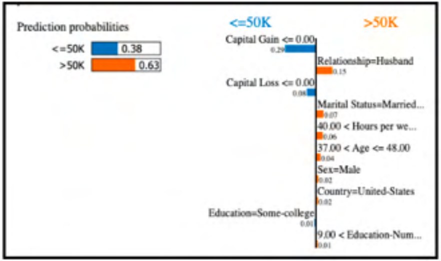


Figure 2: Local explanation in case of Adult dataset (salary prediction)

[3] https://github.com/marcotcr/lime

SHAP: SHapley Additive exPlanationsP[5,6]

Still: an additive feature attribution method, i.e., linear model

$$g(z) = \phi_0 + \sum_{1 \leq i \leq d'} \phi_i z_i,$$

where ϕ_i represents the contribution (importance) of interpretable feature z_i

SHAP: uses Shapley kernel π_x and thus estimation of Shapley values ϕ_i (coalitional game theory) **NB:** KernelSHAP is Costly![4]

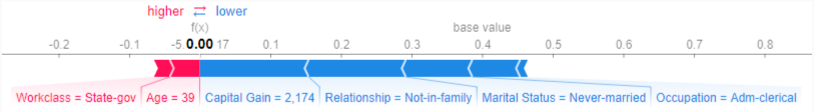


Figure 3: SHAP explanation in case of Adult dataset (salary prediction)

[4] Faster variants like TreeSHAP exist (not model agnostic!)
[5] Lundberg, *et al.* A Unified Approach to Interpreting Model Predictions...
[6] `https://github.com/slundberg/shap`

Tackling unfairness through unawareness: feature dropout and aggregation

Framework to deal Process Fairness

Original Goal: Human-centered approach to reduce a model's dependence on sensitive/salient features while improving its performance

Proposal: Framework consisting of two components:

(i) to assess a model's dependence on sensitive features (fair/unfair)

(ii) (if dependent) to render it fairer (without compromising performance)

Idea: Use a FI-explainer to assess model's dependence sensitive feat.s

Examples: LIME, SHAP and gradient based (under further assumptions)

Here: we focused on model agnostic approaches...

FixOut (FaIrness through eXplanations and feature dropOut)

Fair Model: if its outcomes do not depend on sensitive features

Input: model M, dataset D, sensitive features F, explanation method E

Output: M if fair, otherwise a fairer and more accurate M_{final}

Proposal: **FixOut** with two components

- **$\text{Exp}_{\text{Global}}$:** for global explanations (FI)
- **$\text{Ensemble}_{\text{Out}}$:** Ensemble approach relying on "feature dropout"

FixOut: `https://fixout.loria.fr/`

FixOut (FaIrness through eXplanations and feature dropOut)

Fair Model: if its outcomes do not depend on sensitive features

Input: model M, dataset D, sensitive features F, explanation method E

Output: M if fair, otherwise a fairer and more accurate M_{final}

Proposal: **FixOut** with two components

- **Exp$_{\textbf{Global}}$:** for global explanations (FI)
- **Ensemble$_{\textbf{Out}}$:** Ensemble approach relying on "feature dropout"

FixOut: `https://fixout.loria.fr/`

$\text{Exp}_{\text{Global}}$: model M, dataset D, sensitive F, exp. method E

Idea: Explanations can provide insight into process fairness.

However: LIME and SHAP provide "local" explanations

Solution: Sample a set of instances and aggregate the contributions to estimate the global contribution of each feature.
Example: random or "Sub-modular pick"

Output: k most important (globally) features.

Rule:

If there is **at least one** sensitive feature among the top-k, **then** M is deemed unfair and $\textbf{Ensemble}_{\textbf{Out}}$ applies.

Ensemble$_{\text{Out}}$: model M, dataset D, sensitive features F

Let $a_1, a_2, \ldots, a_k$ be the k features that Exp$_{\text{Global}}$ outputs

Suppose that $a_{j_1}, a_{j_2}, \ldots, a_{j_i}$, $i > 1$, are sensitive (i.e., $\in F$)

Then FixOut trains $i + 1$ classifiers obtained by "feature dropout":

- M_t after removing a_{j_t} from the dataset, for $t = 1, \ldots, i$, and
- M_{i+1} after removing all sensitive features $a_{j_1}, a_{j_2}, \ldots, a_{j_i}$.

Output: Ensemble classifier M_{final} as an aggregation of all M_t's.

Ensemble$_{\text{Out}}$: model M, dataset D, sensitive features F

Example: for an instance x and a class C,

1. **FixOut**: ensemble classifier M_{final} defined as a **simple average**

$$P_{M_{final}}(x \in C) = \sum_{t=1}^{i+1} w_t P_{M_t}(x \in C).$$

2. **FixOut (w)**: Ensemble M_{final} defined as a **weighted average**

$$P_{M_{final}}(x \in C) = \sum_{t=1}^{i+1} w_i P_{M_t}(x \in C),$$

where $w_t = \frac{c_{j_t}}{1+\sum_{u=1}^{i} c_{j_u}}$, $1 \leq t \leq i$, and $w_{i+1} = \frac{1}{1+\sum_{u=1}^{i} c_{j_u}}$ using normalized global feature contributions c_i's.

3. **Alternatively:** use logistic regression (LR) for weight tuning

Example with LIME explanations

FixOut with LIME explanations

Exp$_{\textbf{Global}}$: LIME + random sampling
(of instances and use their explanations to get global explanations)

As before: if Exp$_{\text{Global}}$ outputs $a_1, a_2, \ldots, a_k$ and $a_{j_1}, a_{j_2}, \ldots, a_{j_i} \in F$,
then *FixOut* trains $i+1$ classifiers obtained by "feature dropout":

- M_t after removing a_{j_t} from the dataset, for $t = 1, \ldots, i$, and
- M_{i+1} after removing all sensitive features $a_{j_1}, a_{j_2}, \ldots, a_{j_i}$.

Ensemble$_{\textbf{Out}}$: Ensemble classifier M_{final} defined as

- a simple average (**FixOut**)
- a weighted average (**FixOut (w)**)

FixOut with LIME: RF on German dataset

German Credit Card Score (UCI):

- Applicant profiles (demographic and socio-economic).
- Goal: Predict credit risks (likely & unlikely to pay back)
- Sensitive: 'Statussex', 'telephone', 'foreign worker'

Empirical setting:

- Random Forest: 70% training & 30% test data
- Used: SMOTE oversampling & threshold tuning while training
- Accuracy of M: 0.783

Question: Is this model fair?

FixOut with LIME: RF on German dataset

German Credit Card Score (UCI):

- Applicant profiles (demographic and socio-economic).
- Goal: Predict credit risks (likely & unlikely to pay back)
- Sensitive: 'Statussex', 'telephone', 'foreign worker'

Empirical setting:

- Random Forest: 70% training & 30% test data
- Used: SMOTE oversampling & threshold tuning while training
- Accuracy of M: 0.783

Question: Is this model fair?

FixOut with LIME: RF on German dataset ($\text{Exp}_{\text{Global}}$)

Feature	Contribution
foreignworker	2.664899
otherinstallmentplans	-1.354191
housing	-1.144371
savings	0.984104
property	-0.648104
purpose	-0.415498
existingchecking	0.371415
telephone	0.311451
credithistory	0.263366
duration	-0.223288

Table 1: Top 10 features used by M (by 'submodular pick')

Hence: Model deemed unfair

FixOut with LIME: RF on German dataset (Ensemble$_{\mathbf{Out}}$)

Approach: Train multiple models obtained with feature dropout

- **M1:** Model trained after removing 'foreignworker'.
- **M2:** Model trained after removing 'telephone'.
- **M3:** Model trained after removing the 2 (accuracy of 0.773)
  NB: Accuracy drop when all sensitive features are removed!

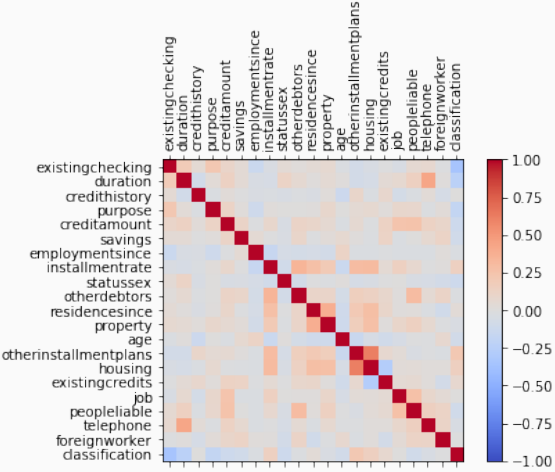
M$_{\mathbf{final}}$: Ensemble of M1, M2 and M3 (accuracy of 0.786)

FixOut with LIME: RF on German dataset

Original		Ensemble	
Feature	**Contribution**	**Feature**	**Contribution**
foreignworker	2.664899	otherinstallmentplans	-1.487604
otherinstallmentplans	-1.354191	housing	-1.089726
housing	-1.144371	savings	0.679195
savings	0.984104	duration	-0.483643
property	-0.648104	**foreignworker**	0.448643
purpose	-0.415498	property	-0.386355
existingchecking	0.371415	credithistory	0.258375
telephone	0.311451	job	-0.252046
credithistory	0.263366	existingchecking	-0.21358
duration	-0.223288	residencesince	-0.138818

Result: M_{final} is “fairer” & at least as accurate (from 0.783 to 0.786)

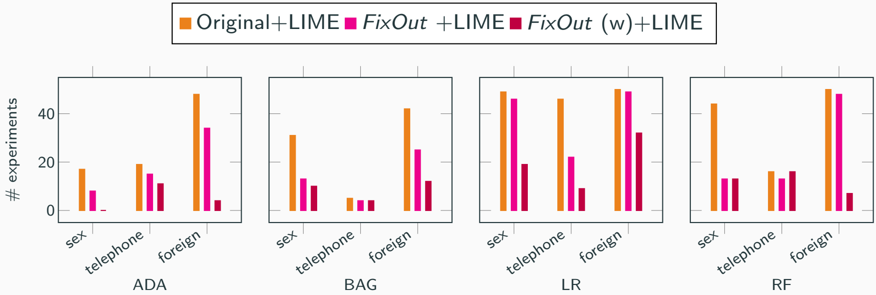
Some preprocessing: What about correlations?



Pearson correlation (German dataset)

Example of available tools: Fairlearn.org

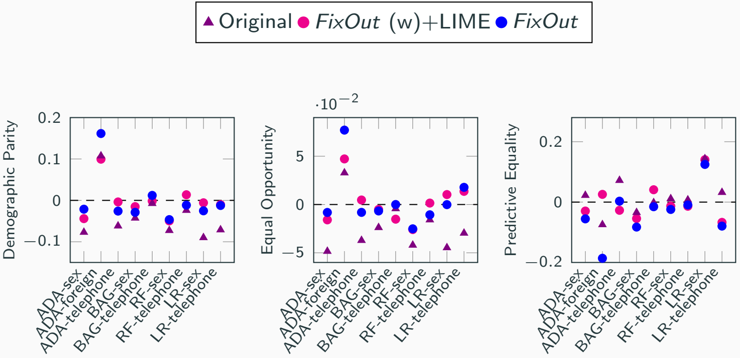
Fairness & Classification assessment (German dataset)



Classification assessment

Dataset	Method	Accuracy				Precision				Recall			
		ADA	BAG	LR	RF	ADA	BAG	LR	RF	ADA	BAG	LR	RF
German	Original	.7362	.7019	.7398	.7556	.5707	.5124	.5716	.6883	.5317	.5738	.5495	.3595
	FixOut	.7419	.7273	.7418	.7598	.5801	.5549	.5754	.7060	.5321	.5371	.5622	.3585
	FixOut (w)	.7405	.7219	.7400	.7583	.5764	.5471	.5708	.7019	.5373	.5076	.5602	.3541

Assessment w.r.t. some fairness metrics (German dataset)



Example with SHAP explanations

FixOut with SHAP: RF on German dataset ($\text{Exp}_{\text{Global}}$)

Same dataset and empirical setting...

Feature	**Contribution**
existingchecking	-7.11624
statussex	-5.950176
housing	-3.27344
job	-2.868195
residencesince	2.832573
telephone	2.290478
property	2.042944
otherinstallmentplans	-1.985275
existingcredits	1.984547
purpose	1.711321

Table 2: Top 10 features used by M

Hence: Model deemed unfair

FixOut with SHAP: RF on German dataset (Ensemble$_{\text{Out}}$)

Approach: Train multiple models obtained with feature dropout

- **M1:** Model trained after removing 'statussex'.
- **M2:** Model trained after removing 'telephone'.
- **M3:** Model trained after removing the 2
  NB: Performance drop when all sensitive features are removed!

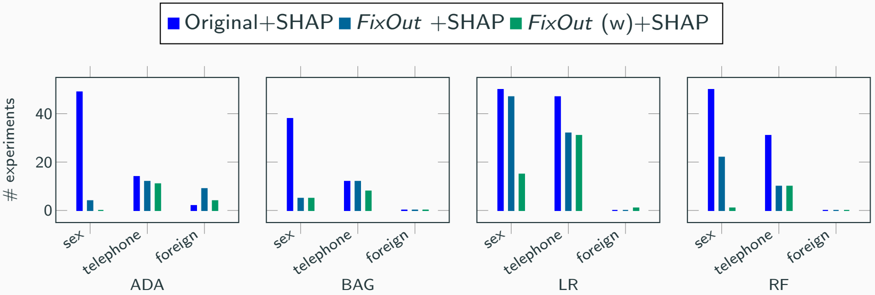
M$_{\text{final}}$: Ensemble of M1, M2 and M3

FixOut with SHAP: RF on German dataset

Original		Ensemble	
Feature	**Contribution**	**Feature**	**Contribution**
existingchecking	-7.11624	existingchecking	-4.285092
statussex	-5.950176	housing	-3.771932
housing	-3.27344	property	3.506007
job	-2.868195	job	-3.061209
residencesince	2.832573	employmentsince	2.646814
telephone	2.290478	existingcredits	2.409782
property	2.042944	otherinstallmentplans	-2.389899
otherinstallmentplans	-1.985275	savings	-2.215407
existingcredits	1.984547	residencesince	2.212183
purpose	1.711321	credithistory	1.188159

Result: M_{final} is fairer & better performance

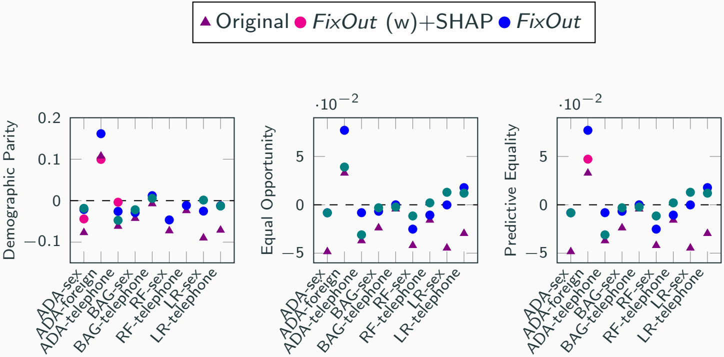
Fairness & Classification assessment (German dataset)



Classification assessment

Dataset	Method	Accuracy				Precision				Recall			
		ADA	BAG	LR	RF	ADA	BAG	LR	RF	ADA	BAG	LR	RF
German	Original	.7362	.7019	.7398	.7556	.5707	.5124	.5716	.6883	.5317	.5738	.5495	.3595
	FixOut	.7419	.7273	.7418	.7598	.5801	.5549	.5754	.7060	.5321	.5371	.5622	.3585
	FixOut (w)	.7427	.7253	.7417	.7613	.5809	.5537	.5746	.7003	.5390	.5142	.5632	.3708

Assessment w.r.t. some fairness metrics (German dataset)



Comparison: Average contribution of sensitive features

No free lunch...

	Method	ADA			BAG			LR			RF		
		sex	telephone	foreign	foreign	telephone	foreign	sex	telephone	foreign	sex	telephone	foreign
German	Original+LIME	-0.13	0.12	3.84	-2.13	0.33	6.36	-13.90	10.08	25.55	-3.29	0.85	23.00
	FixOut +LIME	-0.05	0.09	0.85	-0.63	0.15	1.88	-7.46	2.86	11.90	-0.55	0.67	7.47
	FixOut w+LIME	0.00	0.06	0.02	-0.79	0.11	0.65	-2.00	1.24	3.28	-0.49	0.69	0.23
	Original+SHAP	-0.68	0.10	0.01	-5.13	1.55	0.00	-31.20	11.59	0.00	-10.53	3.21	0.00
	FixOut +SHAP	-0.02	0.08	0.04	-0.76	1.08	0.00	-10.20	3.52	0.00	-1.87	0.69	0.00
	FixOut w+SHAP	-0.07	0.08	0.13	-0.87	0.71	0.00	-1.37	3.25	0.06	-1.87	0.69	0.00

FixOut: brief hands-on

Tutorial & Demo

- **FixOut's start guide** (Jupyter notebook):

  `https://fixout.loria.fr/2020/12/09/tutorials/`

- **Demo**: *FixOut* on selected datasets (tabular data)

  Explanations: LIME, SHAP

  Global explanations : Random Sampling, Submodular-pick

  Aggregation: simple average, weighted average, fine-tuned with LR

  Fairness metrics: demographic parity, equal opportunity, etc.

  `http://vps-9eca9157.vps.ovh.net/`

What about other **data types**?

Example: FixOut on a hate speech classifier

- **Goal**: Classify tweets as *hate speech* or *not*
- **Idea**: Bag of Words (BoW) (**Or:** Groups of words)
- **Dataset**: *Hate speech* dataset [7]

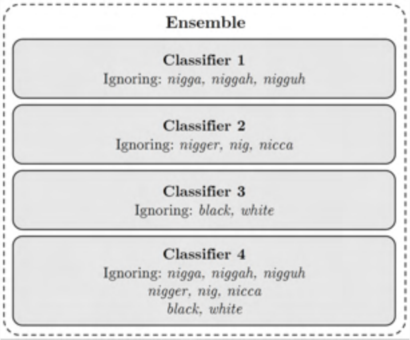


Illustration of textual classifiers used in the ensemble.

[7] Davidson et al. Automated hate speech detection and the problem of offensive language. AAAI. 2017

Textual data: FixOut on a hate speech classifier

Setting: RF classifier, SHAP explanations, RS and BoW

	Without grouping		With grouping	
Word	**Rank**	**Contrib.**	**Rank**	**Contrib.**
niggah	18	0.149	23	0.03
nigger	15	0.164	21	0.031
nigguh	22	0.13	83	0.008
nig	12	0.202	65	0.011
nicca	22	0.107	39	0.018
nigga	20	0.125	12	0.067
white	25	0.087	36	0.018

In fact: Can be used on different data types **e.g.** graphs and other complex data (needs suitable representation...)

Further Resources & Tools

FAT Forensics

- Python toolbox open-sourced for inspecting Fairness, Accountability and Transparency (FAT) aspects of data, models and predictions.
- **b**uild **LIME y**ourself **(bLIMEy)**[8]**:** an algorithmic framework for building custom local surrogate explainers of black-box model predictions, inc. LIME and SHAP
- **Git repository**: `https://github.com/fat-forensics/fat-forensics`

[8] Sokol, *et al.* bLIMEy: Surrogate Prediction Explanations Beyond LIME. *arXiv preprint arXiv:1910.13016*

Other tools

- Fairness assessment (metrics)
- Bias mitigation (e.g. reweighing)
- Visualization

- **IBM AI Fairness 360**[9] (Python, R)
  `https://aif360.mybluemix.net/`
- **Fairlearn** (Python)
  `https://fairlearn.org/`

[9] Bellamy et al. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. 2018. arXiv:1810.01943

Team

Orpailleur[10]


Miguel Couceiro

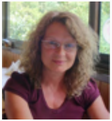

Guilherme Alves


Fabien Bernier

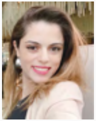

Vaishnavi Bhargava


Amedeo Napoli

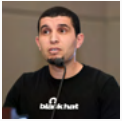
Comète[11]


Catuscia Palamidessi


Ruta Binkyte


Karima Makhlouf


Carlos Pinzon


Sami Zhioua

---

[10] https://orpailleur.loria.fr/

[11] https://team.inria.fr/Comete

Fair and Explainable Models: Feature Issue

EURO J. on Decision Process: Focus on Algorithmic Fairness

Important Dates:

August 31, 2021: Extended abstract

December 15, 2021: Full submission

March 31, 2021: Notification

June 30, 2022: Revision due

Summer 2022: Publication

Call for Paper: Feature Issue on Fair and Explainable Decision Support Systems

Guest Editors:

Miguel Couceiro, University of Lorraine, CNRS, Loria, France (miguel.couceiro@loria.fr)
Luis Galárraga, INRIA Rennes, France (luis.galarraga@inria.fr)

Motivation:

Algorithmic decisions are nowadays being employed on a daily basis. They are carried out by mathematical models trained using machine learning techniques on data collected from past experiences. Well-known examples include decision support systems for loan grants, terrorism detection, prediction of criminal recidivism, and many other activities with social and economic impact on society. While ML-based decision systems generally attain good performance, they can be complex and opaque, not to mention that they are not infallible. This lack of transparency, together with the increasing evidence of biases and unfair outcomes in those systems, has raised several concerns within the scientific and legislative realms.

Most of the notions of fairness focus on the outcomes of the decision process, and they are inspired by several anti-discrimination efforts that aim to ensure that unprivileged groups (e.g. racial minorities) are treated fairly. As such, the problem of improving algorithmic fairness can be formulated as an optimisation one. However, certain dimensions of fairness do not fit into this setting, e.g., fairness through unawareness and counterfactuals. This raises a number of challenges for theorists, researchers, and practitioners.

This brings us to the underlying motivation of this Feature Issue that aims at collecting contributions that focus on the various dimensions of algorithmic fairness, both from foundational and application perspectives. We therefore target works ranging from novel theoretical frameworks to model fairness (and tackle unfairness) in the general case, to the formalization of fairness issues in different applications (from decision making, operations research, resource allocation and policy making) using empirical approaches. Contributions dealing with different data-types, e.g., tabular, sequential, textual and, other complex data such as graphs, are particularly welcome.

Merci de votre attention !

Thank you for your attention!

Grazie mille per la vostra attenzione!

Vielen Dank für Ihre Aufmerksamkeit!

...and let's keep in touch!

Further References

Alves, *et al.* Making ML models fairer through explanations: the case of LimeOut, *AIST'20*.

Bhargava, *et al.* LimeOut: An Ensemble Approach To Improve Process Fairness, *XKDD'20* @ECML-PKDD.

Garreau, *et al.* Explaining the Explainer: A First Theoretical Analysis of LIME, *HCoRR, abs/2001.03447, 2020*.

Grgić-Hlača, *et al.* Beyond distributive fair-ness in algorithmic decision making: Feature selection for procedurally fair learning. *AAAI'18*.

Henin, *et al.* Towards a generic framework for black-box explanations of algorithmic decision systems. *XAI'19* @IJCAI.

Zafar, *et al.* Fairness constraints: Mechanisms for fair classification, *AISTATS'17*.