ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

THEME Fair Learning

Moderators:JM Loubes, L Risser

Université Fédérale
Toulouse Midi-Pyrénées

MEMBERS

Chairs: J.-M Loubes, C. Castets-Renard, F. Gamboa, N. Asher, C. Pagetti, J. Marques-Silva Deel

People involved: (co chairs)

Co-chairs (Loubes): M. Serrurier, B. Laurent

Co-chairs (Castets-Renard): L. Risser

Co-chair (J. Marques-Silva): M. Cooper

Co-Chair (L. Amgoud) : P. Muller

ANITI Resources (post doc, PhD, Mise à Disposition Industrielle, DEEL…)

- 1 Phd ENS T. Benesse (Gamboa-Loubes)
- 1 internship Ecole Polytechnique (04-08/2020)
- MAD DEEL TEAM : Q. Vincenot (Thales), J. Sen Gupta (Airbus), A Martin-Picard (Scalian), D. Vigouroux (IRT DEEL)
- MAD (C. Pagetti chair) M. Belcaid (CS group)
- 2 PhD (C. Castets-Renard chair) Ronan Pons + Evgeniia Volkova (CIFRE Toulouse Métropole)

Other Resources (other Phd project)…

- 1 Phd CIMI P. Gordaliza (Gamboa-Loubes-del Barrio)
- 1 Phd X L. De Lara (Asher, Loubes, Risser)
- 1 Phd Region C. Champion (Burcelin-Loubes)
- 1 Post Doc Valladolid H. Hinouzhe (del Barrio, Loubes)
- 2 Internships Ecole Polytechnique (Loubes/Risser)
- 2 internships Univ. Toulouse 3 (Loubes/Risser)

ANR Projects :

- ANR SLANT (Asher/T. Muller)

Definition/scientific perimeter of the theme

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE

Thread 1: Fair Learning : Analysis of Bias for Fairness

Thread 2: Analysis of Bias for Data and algorithms in critical system design and certification

Thread: Fair Learning

Define methods based on Machine Learning and hybrid AI for detecting, controlling and removing unwanted biases in Machine Learning.

Objective: This thread deals with novel methods to detect and then to limit undesired biases made by AI systems that end users could experience. This thread also investigates how these methods respond to legal and ethical requirements for AI systems.

Challenges:

1. Provide **formal and legal definitions of biases** that can lead to **tractable and feasible** controls over the algorithm and also scale well to large volumes of data to be used in real-world applications.
2. Understand **the nature & epistemological consequences of bias** (distribution of the learning sample, sampling of the dataset, legal or technical constraints)
3. Understand the **effect of fairness** conditions on the performance of the AI system.
4. Using **bias for explainability** counterfactual & logical methods.

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

Tools/Techniques:

- Statistical modeling, Machine Learning and Neural networks
- Logic and Semantic Modeling
- Optimal Transport Theory, Theory of Deformations
- Counterfactual Reasoning

*A*pplication/Use Cases:

- *Images (CelebA : Algorithmic bias when predicting whether someone is attractive based on >200K RGB pictures)*
- *Quantitative Data (Adult Census Income : Predict whether income exeeds $50K/yr based on census data)*
- *NLP Bias (RH and Recruitment in Bios : NLP for the prediction of different features based on >1500K bias in bios, twitter accounts and newspapers, “cockpit - ATC discussions” Airbus data set)*

Highlight & main results

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE

Publications

1. **A survey of bias in Machine Learning through the prism of Statistical Parity for the Adult Data Set**. P Besse, E del Barrio, P Gordaliza, **JM Loubes, L Risser.** To appear in American Statistician 2021.
2. **Obtaining fairness using optimal transport theory**.P Gordaliza, E Del Barrio, G Fabrice, **JM Loubes**.International Conference on Machine Learning, 2357-2365, 2020.
3. **Can Everyday AI be Ethical? Machine Learning Algorithm Fairness,** P Besse, **C Castets-Renard,** A Garivier, J**M Loubes**. Statistiques et Société (vol 6), 2019
4. Comment construire une intelligence artificielle responsable et inclusive?" (2020), C. Castets-Renard, Recueil Dalloz
5. Accountability of Algorithms: A European Legal Framework on Automated Decision- Making (2019), 30 Fordham Intell. Prop. Media & Ent. L.J. 91, C. Castets-Renard
6. **Review of Mathematical frameworks for Fairness in Machine Learning**. E. Del Barrio, P. Gordaliza, **J.-M. Loubes**. ArXiv preprint arXiv:2005.13755, 2020, submitted to Statistical Science.
7. **Bias in Semantic and Discourse Interpretation. N. Asher,** S. Paul, *Linguistics and Philosophy (*minor corrections*),* 2020
8. **Interpretive blindness and the impossibility of learning from testimony**. **N. Asher**, J. Hunter. submitted to *Coling 2020.*
9. **Fair and Adequate Explanations**. **N. Asher,** S. Paul, C. Russell. ArXiv preprint:2001.07578, 2020
10. **Towards Formal Fairness in Machine Learning**. A. Ignatiev, M. C. Cooper, M. Siala, E. Hebrard, **J. Marques-Silva**:. *CP* 2020: 846-867
11. **Projection to Fairness in Statistical Learning.** Thibaut Le Gouic, **Jean-Michel Loubes,** Philippe Rigollet,(submitted)
12. **Entropic Variable Projection for Model Explainability and Interpretability.** F. Bachoc, F. **Gamboa**, M. Halford, **J.M. Loubes, L. Risser**. Arxiv https://arxiv.org/abs/1810.07924, 2020

Highlight & main results

Scientific event organization (conference, workshop, GDR) & participation

2019 : GDR MaDICS CNRS J-M. Loubes (Rennes 2019 tutorial on fairness in Machine Learning), Journées de la Statistique Francaise J-M. Loubes (Nancy Cours Fair Machine Learning), 2019 (ICML Oral Presentation), Meetings of Oberwolfach (Oral Presentation)
2020 : L. Risser: Co-organisation with E. Gondet (CNRS/OMP) of the parallel sessions of JDEV 2020 dealing with AI (>150 people).
2020: Organisation of the ICML workshop **Law and Machine Learning**. C. Castets-Renard (U. Ottawa/Law), S. Cussat-Blanc (U. Toulouse/IRIT), L. Risser (CNRS/applied maths). About 50 people.
2020: Organisation of Workshop India Kolkatta (Explainable and Fair Machine Learning) J-M. Loubes

Dissemination (main)

J-M. Loubes Nuit Européenne de la Recherche (09/2019)

J-M. Loubes 'Fairness and Risk' Singapore Seminar for Acturials (08/2020)

J-M. Loubes 'Fairness' IDIAP Research Institute (04/2020)

N. Asher 'Fair and Adequate Explanations,' Workshop CIFAR/CNRS "Fairness, Interpretability and Privacy for Algorithmic Systems", 3-4 June, 2019, Turing Institute London, UK.

L. Risser and R. Pons (PhD student ANITI) 'Explainability and Fairness of Black-box decision rules in AI'. Meetup Machine Learning Pau 03/2020

L. Risser and R. Pons (PhD student ANITI) 'Explainability and Fairness of Black-box decision rules in AI'. CNRS/DEVLOG workshop APSEM. 2019.

Thread: Analysis of Bias for Data and algorithms in critical system design and certification

Objective: Developing and applying fairness techniques for bias found in industrial applications in order to support certifiable AI.

In industrial applications, bias comes from:

- unbalanced representation of operating and environmental conditions,
- wrong labelling
- incomplete description of data leading to spurious correlations

This bias is problematic if it is not the same on the test distribution (usage of the model).
An industrial model must be **robust to changes** in the bias => **Robustness**

Challenges:

- To understand the effects of the distribution of the learning sample in the machine learning process and the generalization error in order to **guarantee performance** of the algorithm **resilient to modifications of their environment (transfer learning, consensus learning, protection from adversarial conditions)**
- To **detect unspecified bias** that may hamper the machine learning system.

Tools/Techniques:

- Robustness, Distributional Shift and Consensus Learning
- Uncertainty quantification and Sensivity Analysis
- Unsupervised Method & Clustering
- Logic and compliance to formal rules

Application/Use Cases:

- Applications to Industrial Use Case : Images Blink Renault data set, Anomaly of Electronic Components (Airbus) Satellite Images (Thales) Breaking distance estimation (numeric Airbus)
- Applications to Medical Data (cytometry data, genomic data set on hepatitis and diabetes fat liver)

Highlight & main results

Publications

1:**A central limit theorem for L_p transportation cost on the real line with application to fairness assessment in machine learning**.E Del Barrio, P Gordaliza, JM Loubes.Information and Inference: A Journal of the IMA 8 (4), 817-849, 2020.
2:**optimalFlow: Optimal-transport approach to flow cytometry gating and population matching.** Eustasio del Barrio, Hristo Inouzhe, Jean-Michel Loubes, Carlos Matrán, Agustín Mayo-Íscar. To appear in BMC-Bioinformatics 2021.
3:: **Tackling Algorithmic Bias in Neural-Network Classifiers using Wasserstein-2 Regularization**. Risser L., Vincenot Q., Loubes J.M.Arxiv 2020 (submitted)
4: **Sobol indexes : new methods to quantify causal fairness.** C. Benesse, F. Gamboa, J-M. Loubes (preprint)
5: **Explaining black box models through optimal stressing (**Medium) A. Gauffriau A. Picard

Python Package for Biased Dataset generation towards robustness, explainability and causality : gems-ai.com

Scientific event organization (conference, workshop, GDR) & participation

Mobility AI (2019), ICIAM Valence (2019), P-IJCAI (2019)

Where we go ? Journey to Fair(y) Land

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

Theme roadmap (year 2, 3 and 4 - cf p38 roadmap)

1: find real case of industrial /medical bias

2: detect “unknown uncategorized” biases

3: hybrid methods for bias ---e.g., equivalences between deformational and logical approaches to Counterfactual reasoning.

4: Applications to NLP

5: Investigate interactions between robust learning and methods for detecting and removing bias

6: from Legal constraints to tractable constraints for the algorithm (and vice & versa)

How we go ? On-going work

On-going collaboration between chairs

1: Loubes/Risser/Gamboa/Deel Bias for explainability

2: Loubes/Risser/Gamboa/Deel (+phd) Fairness & Sensitivity Analysis

3: Asher/Castets-Renard(Risser)/Loubes (+phd) counterfactual reasoning

4: L. Amgoud chair/J. Marques-Silva chair logic based for fairness

On-going collaboration with ANITI Industrial partners

1 industrial working group Loubes/Risser/J. Sen Gupta with

Airbus/Thales/IRT Saint-Exupery, Airbus, Renault, Thales, Scalian detection of unknown bias

With whom

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE

On-going collaboration with external projects (national, EU, industry…)

- **INSERM**
- **INRIA Magnet Team Lille**

- **University of Valladolid,**
- **University of Luxemburg,**
- **Basque Center for Applied Mathematics,**

- **University of Ottawa,**
- **MIT Boston,**
- **Monash University Australia**
- **MILA Montreal.**

- **EDF**

On the way ...

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

I. **Emerging collaboration between chairs**

1 Castets-Renard/Loubes from law to science and vice-versa

2: Travé/CS-Group (M. Allain-Moutet)/IRT/Loubes Bias from unbalanced data set in anomaly detection

II. **Future collaborations to be developed**

Applications to medical data sets and medical database

Scientific animation of the theme

Monthly Working Group Meetings (open Aniti wide)