ANITI

ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

Explainability

E. Lorini & J. Marques-Silva

September 28-29 2020

Université Fédérale
Toulouse Midi-Pyrénées

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

Members

Theme Description

Ongoing Work

Highlights & Main Results

Scientific Animation

Members I

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

Chairs

- T4.C1 L. Amgoud: Argumentation
- T4.C2 J.-M. Loubes: Fair and Robust Learning (FRL)
- T4.C3 J. Marques-Silva: Deep Learner Explanation & Verification (DeepLEVER)
- T4.C4 T. Schiex: Design using intuition[1] and logic[2] (DUIL)
- T4.C5 L. Travé-Massuyès: Synergistic transformations in model based and data based diagnosis (SynT)
- T4.C6 D. Vigoroux: DEEL

Others?

Members II

Co-Chairs (in order)

- T4.C1 E. Lorini mostly (AR for planning, expl. using epistemic logic, SAT, QBF), P. Muller (Language progressing)
- T4.C2 B. Laurent, M. Serrurier
- T4.C3 M. Cooper (IRIT), E. Hébrard (LAAS)
- T4.C4 Sophie Barbe (INSA/INRAE), David Simoncini (IRIT), Georges Katsirelos & Simon de Givry (INRAE)
- T4.C5 Nathalie Barbosa Roa (Vitesco Technologies), Elodie Chantery (LAAS), Xavier Pucel (ONERA)

Members III

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

Associated researchers

- T4.C1 Andreas Herzig, Frederic Maris and Dominique Longin
- T4.C3 Mohamed Siala (LAAS)
- T4.C4 David Allouche, Nathalie Rousse (INRAE)
- T4.C5 Yannick Pencolé (LAAS), Gregor Gössler (INRIA-RA) , Thomas Mari (PhD, INRIA-RA),Stéphanie Roussel (Onera)
- T4.C6 Edouard Pauwels, Jean-Michel Loubes, Thomas Serre

Members IV

PhD students and Post-docs

- T4.C1 Louis Rivière, Tom Portoleau* (PhD), Nicolas Schmidt (Post-doc) (PhD), Henri Trenquier (PhD), Vivien Beuselinck (PhD), Xinghan Liu (PhD)
- T4.C3 Yacine Izza (Post-doc), Thomas Gerspacher* (PhD), Xuanxiang Huang (PhD)*
- T4.C5 Valentin Bouziat* (PhD)
- T4.C6 Thomas Fel (CIFRE SNCF)

* Not funded by ANITI

Members V

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

MAD (mise à disposition/industrial)

T4.C6 Florence DE GRANCEY (Thales), Mikaël CAPELLE (IRT Saint Exupery), Adrien GAUFFRIAU (Airbus), Agustin MARTIN PICARD (Scalian), Mélanie DUCOFFE (Airbus), Raphael PUGET (Renault), Frederic BOISNARD (Renault), Bertrand CAYSSIOLS (Renault), David VIGOUROUX (IRT Saint Exupery), Nathalie BARBOSA (Vitesco Technologies)

Theme Description I

Logic vs Statistics for Explainability

IP	Acceptable AI + Certifiable AI
titre	4 Explainability

Description

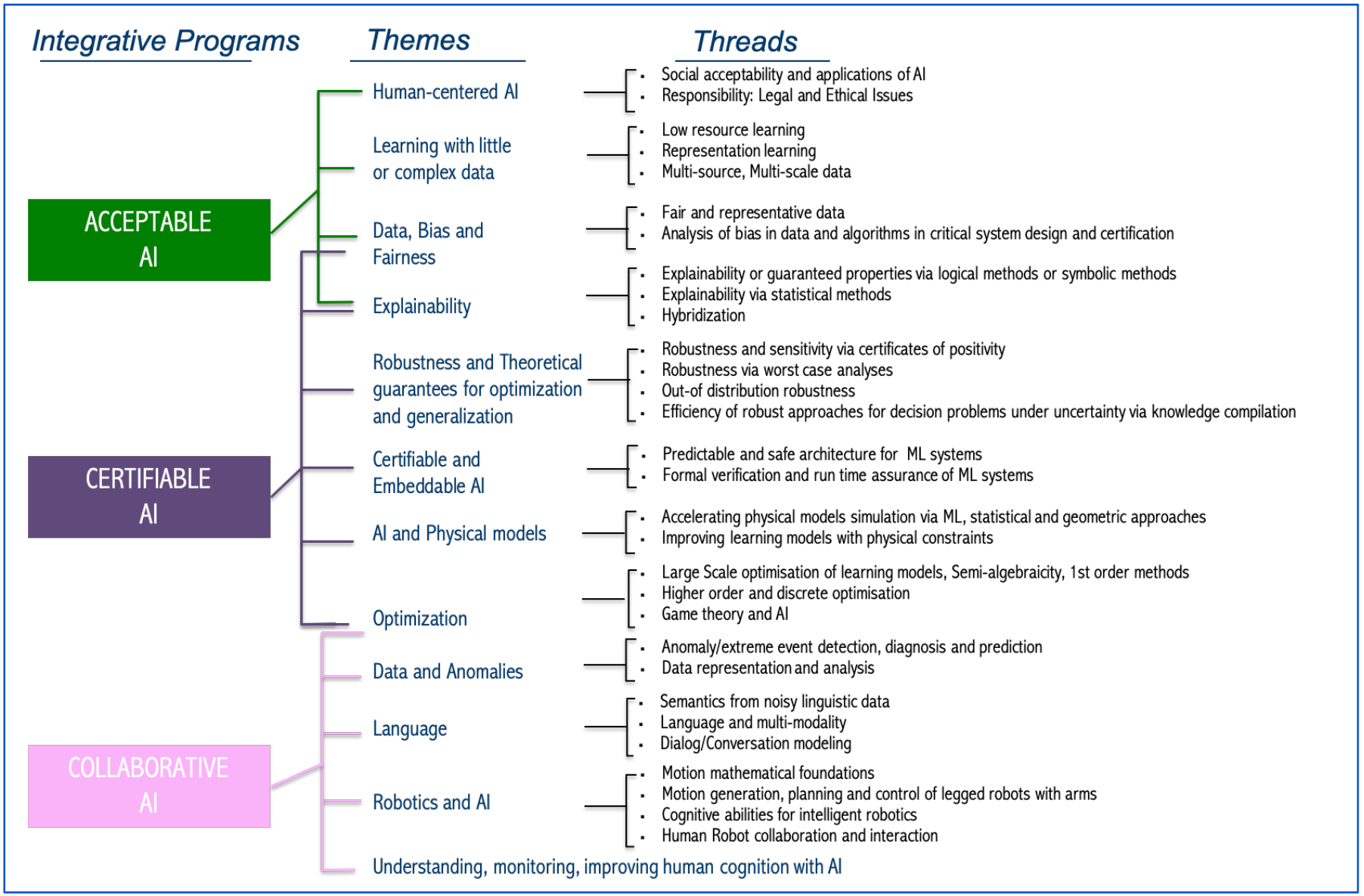
As machine learning systems have become more and more complex, they have managed to achieve performances in some areas that surpass human ones. However, it is often difficult to impossible to determine the reasons behind the predictions of a complex system. Thus explainability or the interpretability of an ML system has become an important issue in the foundations of AI. And explainability is a core theme of both Acceptable and Certifiable AI in ANITI. Both IPs will monitor this theme, though researchers from Collaborative AI (Schiex, Travé) will also participate.
This theme currently features three threads, one featuring a logical approach to explainability, the second a statistical approach, and the third one addresses hybridization approaches that aim to leverage the best of both techniques.

Threads

4.1 Explainability or guaranteed properties via logical methods or symbolic methods
4.2 Explainability via statistical methods
4.3 Hybridization

Theme Description II

Logic vs Statistics for Explainability



Theme Description III

Logic vs Statistics for Explainability

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

Thread 4.1: Explainability with logical methods

Devise logical-based approach to represent and reason about explanations;
Approaches can be blackbox-based or whitebox-based.
Technologies to exploit: Oracles for NP/PSPACE, e.g. SAT, SMT, MILP;
Optimization: MaxSAT, QMaxSAT, etc.; (approximate) model counting.

Thread 4.2: Explainability with statistical methods

Translate some of the insights from the logical method into a more mathematical framework employing notions from geometry and topology to provide a hybrid approach to explanation. This approach may offer a way to overcome the limitations of the logical method. Automated reasoning techniques for computing explanations automatically will be limited to small learning networks; in principle the mathematical representations will scale up better.

Thread 4.3: Hybrid XAI – combine logical & statistical methods

To devise methods integrating logic (rigor) and statistics (efficiency) that will represent the next generation of XAI tools

Ongoing Work I

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

T4.C1 Chair: Argumentation
Black-box ML models: Axiomatic theory of explanation (properties, evaluation methods); Formalizing existing types of explanation in a unified setting; Dialectical explanations; Argument-based classifiers.
Endowing agents with explanatory capabilities: an approach based on epistemic logic; A modal language for representing explanations and biases in classifier systems

T4.C2 Chair: FRL
Entropic Variable Boosting

T4.C3 Chair: DeepLEVER
Abductive explanations; Logical representations of ML models; Tractable explanations; Assessment of heuristic explanations; Principled heuristic explanations; Links with fairness; Links with interpretability

Ongoing Work II

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

T4.C4 Chair: DUIL
Learning graphical models – learn to reason

T4.C5 Chair: SynT
Providing explanations about the behavior of a system modelled in a discrete event framework (automata). Explanations may be of type 1) "What did happen? (diagnosis)" or of type 2) "Why did this happen? (explanation of property violations)". Also, some work on active diagnosis (which actions to refine diagnosis?).

T4.C6 Chair: DEEL
Overview of state-of-the-art written; exploring two research themes: "measure to quantify the reliability of an explaination" and "application of formal methods to a given explaination". Evaluation of three technologies "Internal model analysis", "Building features/attentions models/ Unsupervised learning for representation disentanglement" and "formal methods"

Collaborations & Grants

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

Ongoing Collaborations

1. ML2R, Univ. Dortmund
2. IRISA, Univ. Rennes
3. CRIL, Univ. Artois
4. Monash University
5. Univ. Singapore (ANR/NRF)
6. VMWare Research
7. Politecnico di Torino
8. ...

Grants

1. ICT-38 Coala
2. Several proposals under review

Publication Highlights I

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

T4.C1 Chair: Argumentation

1. KR'20:
   "Explaining Black-box Classification Models with Arguments"
2. Submitted 2020:
   "Interpretable Embeddings: a Simple but Effective Means for Bias Detection in NLP"
   "A Formal Comparison of Various Types of Explanation"
   "A Computationally Grounded Logic of Graded Belief and its Application to Explanation Modeling"
   "Dialectical explanations of classifiers"

T4.C2 Chair: FRL

1. CoRR'18:
   "Entropic Variable Boosting for Explainability and Interpretability in Machine Learning"

Publication Highlights II

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

T4.C3 Chair: DeepLEVER

1. AAAI'19:
   "Abduction-Based Explanations for Machine Learning Models"
2. NeurIPS'19:
   "On Relating Explanations and Adversarial Examples"
3. SAT'19:
   "Assessing Heuristic Machine Learning Explanations with Model Counting"
4. NeurIPS'20:
   "Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay"
5. CP'20:
   "Towards Formal Fairness in Machine Learning"
6. IJCAI'20:
   "Learning Optimal Decision Trees with MaxSAT and its Integration in AdaBoost"
7. CoRR'20: "MurTree: Optimal Classification Trees via Dynamic Programming and Search"

Publication Highlights III

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

T4.C5 Chair: SynT

1. DX'19:
   Towards Causal Explanations of Property Violations in Discrete Event Systems
2. AAMAS'19:
   Preference-Based Fault Estimation in Autonomous Robots : Incompleteness and Meta-Diagnosis

T4.C6 Chair: DEEL

1. Preprint, Sep'20:
   Representativity and Consistency Measures for Deep Neural Network Explanation

Dissemination & Animation

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

- Program (co)chair: CP 2019, CPAIOR 2020, "AAAI Sister Conferences track" at AAAI2020
- Area Chair: IJCAI 2020 (x2), KR 2020, IJCAI 2019
- Tutorial chair: IJCAI 2020
- Tutorials AAAI 2020, IJCAI 2020, PFIA 2019, CP 2020, STACS 2020
- Invited talks KIM Data and Life Science 2019, JOBIM 2019, DATARMOR'2020, ICLP 2020, ...
- Organization of the ACP / GDR IA / GDR RO summer school 2020 : *Combinatorial optimization, constraint programming and machine learning*
- Organisation Project Management and Scheduling (PMS) 2020 (postponed to 2021)
- Editorial boards: AIJ Associate Editor

Proposed Animation Plan

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

1. Theme 04 regular workshop
   - Presentation-led
   - Focus on brainstorming about inter-chair/intra-theme research
2. Identify inter-chair/intra-theme/inter-theme challenges
3. Intra-theme collaborations
4. Inter-theme collaborations
5. Promote industry collaborations
6. Exploit industry feedback
7. Promote joint supervisions

Intra/Inter Theme Collababorations

ongoing or potential

Intra theme collaborations

- L.A.+J.-M.Loubes: collaboration on metrics/properties of explanations
- L.T.+JMS: diagnosis

Inter theme collaborations

- Language theme

Industrial Partners

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE
TOULOUSE INSTITUTE

T4.C3 Chair: DeepLEVER

Caisse D'Epargne and ADAGOS
Thales: initial contact

T4.C5 Chair: SynT

Vitesco Technologies; Nukkai

T4.C6 Chair: DEEL

Renault Software; SNCF; Thales; Airbus; Continental; Safran; Scalian; etc.