RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

cea cnrs Inria

FRANCE 2030 PROGRAMME DE RECHERCHE *INTELLIGENCE ARTIFICIELLE*

SAIF: Safe AI through Formal Methods

Caterina URBAN (ANTIQUE, Inria Paris)
Zakaria CHIHANI (LSL, CEA-List)

anr agence nationale de la recherche

Project Members

Caterina URBAN

ANTIQUE

Inria Paris

Nathanaël FIJALKOW

LaBRI

Université de Bordeaux

Sylvie PUTOT

LIX

École Polytechnique

Benedikt BOLLIG

LMF

Université Paris-Saclay

Zakaria CHIHANI

LSL

CEA-List

Eric FABRE

SuMo

Inria Rennes

Guillaume CHARPIAT

TAU

Inria Saclay

Context and Motivation

Machine Learning in High-Stakes Systems

data

ML software

perform tasks that are impossible using explicit programming

act as surrogate model

automate decision-making

Context and Motivation

Challenges

1. ML-based systems are **LARGE**
   They can have hundreds of millions of parameters;
2. ML-based systems are **DIFFICULT TO SPECIFY**
   They are built from large example bases, rather than from well-structured specifications;
3. ML-based systems are **MONOLITHIC**
   They are rarely decomposable in smaller components each with its own specification;
4. ML-based systems are **OPAQUE**
   They are prone to bias, they are difficult to interpret, …
5. ML-based systems are **HETEROGENEOUS**
   They employ different architectures, different activation functions, …

Project Goal

Harness and *rethink* decades of work in **FORMAL METHODS**
to tackle the modern challenges of MACHINE LEARNING

Scientific Objectives

1. **Specifying ML-based Systems**

   Pushing the boundaries in the landscape of **formal specification techniques for ML-based systems**

   Motivated by challenge 2 (difficult to specify) and challenge 3 (monolithic)

2. **Validating ML-based Systems**

   Tackling the verification of a **much broader spectrum of properties** for a **much more comprehensive spectrum of ML-based systems**

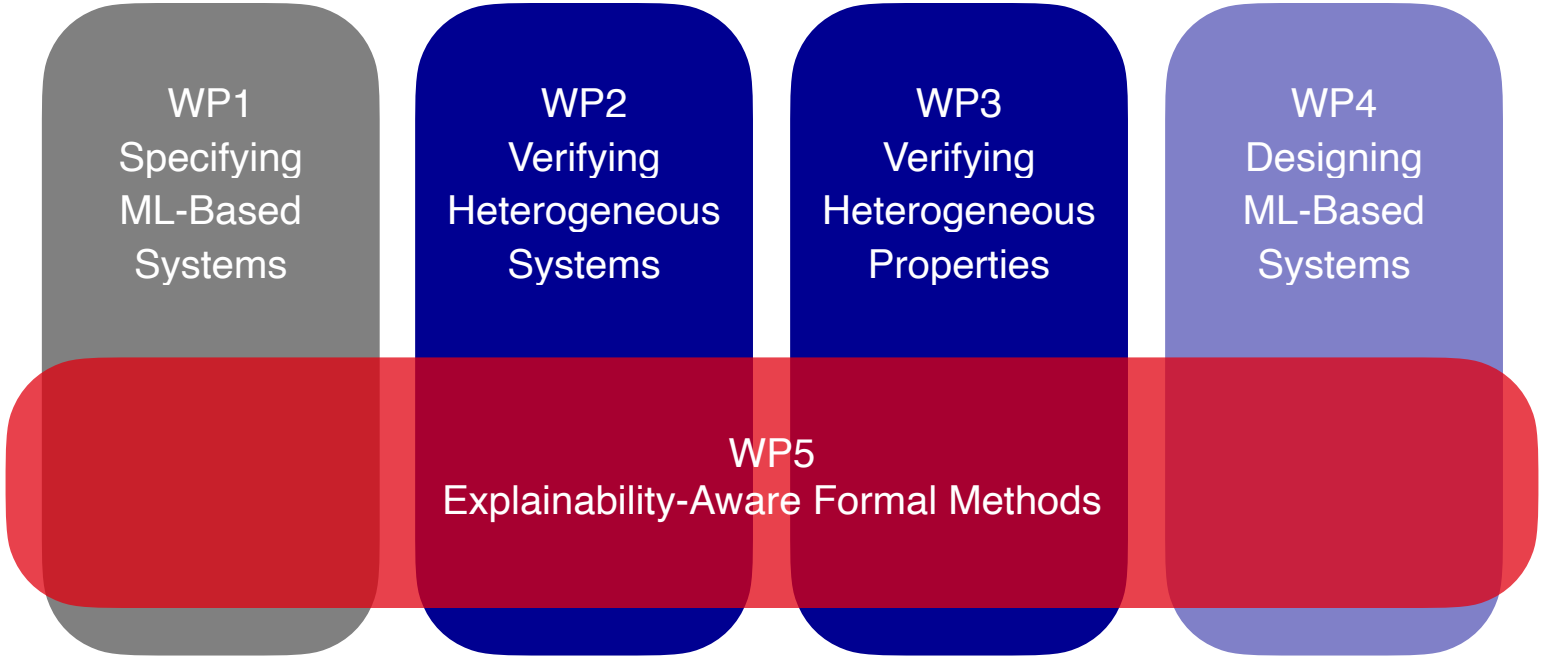
   Motivated by challenge 1 (large) and challenge 5 (heterogeneous)

3. **Guiding the Design of ML-based Systems**

   Making formal methods an asset in the design of more trustworthy and easier to verify ML-based systems

   Motivated by challenge 2 (difficult to specify), challenge 3 (monolithic), and challenge 4 (opaque)

Action Plan


WP1
Specifying
ML-Based
Systems
WP2
Verifying
Heterogeneous
Systems
WP3
Verifying
Heterogeneous
Properties
WP4
Designing
ML-Based
Systems
WP5
Explainability-Aware Formal Methods

WP1: Specifying ML-based Systems

LaBRI

specification mining for temporal properties

LIX

verification of temporal logic properties cyber-physical systems

WP1 Specifying ML-Based Systems

- **Principled Synthetic Data Generation**
  - **1 PhDs (CEA+TAU)**
  - 1 PhD (ANTIQUE)
  - 1 Postdoc (LaBRI)
- **Synthesis of Temporal Logic Specifications**
  - 1 PhDs (LIX)
  - 1 Postdoc (LaBRI)

WP2: Verifying Heterogeneous Systems

ANTIQUE

static analysis and certified training of ML software

SuMo

verification and control of large scale systems

WP2
Verifying
Heterogeneous
Systems

- **Open, Modular, Unifying Verification Framework**
  - 2 PhDs (CEA)
  - **1 PhD (SuMo+ANTIQUE)**
- **Dynamical Systems**
  - **1 PhDs (SuMo+LaBRI)**
  - 1 Postdoc (SuMo)
- **Advanced Neural Network Architectures**
  - **1 PhDs (ANTIQUE + CEA)**
  - 1 Postdoc (LIX)

WP3: Verifying Heterogeneous Properties

LIX

set-based methods and numerical systems analysis

LaBRI

generator-based global robustness verification

WP3
Verifying
Heterogeneous
Properties

- **Beyond Local Classification Robustness**
  - 2 PhDs (ANTIQUE)
  - 1 PhD (LIX)
  - 1 Postdoc (ANTIQUE)
- **Probabilistic Properties and Hyperproperties**
  - 2 PhDs (LIX)
- **Generator-Based Properties**
  - **1 PhDs (CEA+TAU)**
  - 1 PhD (LaBRI)

WP4: Designing ML-Based Systems

LSL

formal methods for ML-based systems in industrial settings

LMF

verification of infinite-state, distributed, hybrid, and stochastic systems

WP4 Designing ML-Based Systems

- **Monitoring, Harnesses, and Fail-Safe Procedures**
  - 1 PhDs (ANTIQUE)
  - 1 PhD (LIX)
- **Principled Training Approaches**
  - 1 PhDs (ANTIQUE)
  - 1 Postdoc (LaBRI)
  - 1 Postdoc (ANTIQUE)
- **Reinforcement Learning**
  - 1 Postdoc (LMF)

WP5: Explainability-Aware Formal Methods

ANTIQUE
static analysis

TAU
deep learning

LSL

formal methods for ML-based systems in industrial settings

LaBRI

reinforcement learning and program synthesis

WP5 Explainability-Aware Formal Methods

- **Verification for Explainability and Explainability for Verification**
  - 1 PhDs (ANTIQUE)
  - **1 PhD (TAU+CEA)**
  - 1 Postdoc (ANTIQUE)
  - 1 SRP (TAU)
- **Case-Based Reasoning**
  - 1 PhDs (CEA)
- **Explainable Reinforcement Learning**
  - **1 PhD (LaBRI+SuMo)**
  - **1 Postdoc (LaBRI+SuMo)**
  - 1 Postdoc (SuMo)

Expected Outcomes

Scientific Outcomes

1. Publication of scientific papers in top-tier journals and conferences in formal methods and machine learning
2. Organization of seminars, workshops, and summer schools

Solid theoretical foundations for continuing our quest for trustworthy ML

2. Technological Outcomes

1. Release of openly-available libraries of code to help future developments
2. Creation and release of benchmarks to evaluate and compare methodologies and tools

Solid practical foundations for continuing our quest for trustworthy ML

3. Societal Outcomes

1. New specification frameworks for ML trustworthiness properties
2. New methods and tools to characterize the validity of these properties
3. New design guidelines

Support for certification processes of ethical, transparency, safety, and security standards

Improvements of overall **quality and reliability of ML-based systems**

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

cea cnrs Inria

FRANCE 2030

PROGRAMME DE RECHERCHE

INTELLIGENCE ARTIFICIELLE

Retrouvez toutes nos actualités

anr agence nationale de la recherche