# Dependable ans Trustworthy AI in Trieste

**Luca Bortolussi , Francesca Cairoli , Ginevra Carbone , Andrea De Lorenzo , Eric Medvet , Laura Nenzi**

## 1 Introduction

The activity in the area of dependable and trustworthy AI in the University of Trieste node of the AIIS lab is mainly concentrated in two departments (Mathematics and Geosciences and Engineering and Architecture), specifically in two research Labs:

- Machine Learning Lab[1]
- Artificial Intelligence and Cyber-Physical Systems Lab

Currently, more than ten researchers are involved in the projects, including Full Professors, Associate Professors, Tenure-track researchers, Post-Docs, and PhD students. These two labs and the current projects are described below.

## 2 Research Groups

### 2.1 Machine Learning Lab

The Machine Learning Lab (MaLe Lab) is led by Prof. Alberto Bartoli and Prof. Eric Medvet and deals with Machine Learning (ML) applications to problems of engineering interest.

The research carried out in the lab exploits a broad toolbox of techniques, including classical ML, Deep Learning (DL), Reinforcement Learning (RL), and Evolutionary Computation (EC). Those techniques are applied to different domains and for different purposes, including cybersecurity [Canfora *et al.*, 2015; De Lorenzo *et al.*, 2020], information extraction [Bartoli *et al.*, 2017; Bartoli *et al.*, 2019], image, signals, and text generation [Castelli *et al.*, 2021; Bartoli *et al.*, 2016], in particular using Generative Adversarial Networks (GANs), synthesis of policies and rules for cyber-physical systems (CPSs) [Talamini *et al.*, 2020; Pigozzi *et al.*, 2021], and robot control and automatic design (jointly with the Evolutionary Robotics and Artificial Life lab[2], led by Prof. Eric Medvet) [Talamini *et al.*, 2021; Salvato *et al.*, 2021]. In particular in the EC field, the lab research also consists in theoretical and algorithmic contributions [Manzoni *et al.*, 2020; Bartoli *et al.*, 2018].

[1] `https://machinelearning.inginf.units.it/`

[2] `https://erallab.inginf.units.it/`

### 2.2 Artificial Intelligence and Cyber-Physical Systems Lab

The Artificial Intelligence and Cyber Physical Systems Lab (AI-CPS) is led by Prof. Luca Bortolussi, and develops novel approaches combining ML with symbolic techniques and quantitative formal method approaches to improve efficiency and explainability/dependability of ML systems.

The lab is currently focusing on several directions, mainly:

- neuro-symbolic computing, integrating ML with logic to improve learning of logical classification rules for temporal signals;
- robust and explainable deep learning, investigating adversarial robustness of prediction and explanations in the context of Bayesian methods;
- applications of ML techniques for safe controller synthesis and monitoring of CPSs.

## 3 Select current research projects

### 3.1 Metaheuristics for interpretable AI

In collaboration with the Life Sciences and Health Group of the Centrum Wiskunde & Informatica (CWI) of Amsterdam, we are exploring metaheuristic approaches to develop AI models which are easily interpretable by human users.

Nowdays, many risk-sensitive applications require AI models to be interpretable, particularly when these have a strong impact on people. Interpretable models are important as a safety measure for real-world applications, as a tool for detecting bias, and as a way for increasing social acceptance of AI. Current researches are focused on tuning, by trial-and-error, hyper-parameters of model complexity that are only loosely related to interpretability. Our research efforts are aimed at the development of a meta-learning approach: we have shown how a non-trivial proxy for interpetrability can be learned using a ML approach. At the moment, we have shown how this approach can be applied to evolutionary symbolic regression by validating our results on five synthetic and eight real-world symbolic regression problems and comparing

to the traditional use of solution size minimization [Virgolin *et al.*, 2020].

Moreover, another problem related to interpretability is its intrinsically subjectivity: whether a person finds a model interpretable depends on that person's background. Within the same collaboration, we are investigating the problem of capturing the subjective notion of interpretability using active learning. Instead of attempting to infer a general model for interpretability, we train an estimator tailored to the user's perception of what is interpretable and what is not. In our recent work [Virgolin *et al.*, 2021], we have shown how—only asking the user to tell, given two models at the time, which one is less complex—it is possible to learn an interpretability estimator that can be very different for different users. Our experiments, which involved a large number of participants, showed how a personalized interpretability model is possible with a human-in-the-loop approach and also highlighted how the users tend to prefer personalized models over traditional interpretability models.

### 3.2 Combining ML and logic for robust and explainable classification of temporal signals

ML techniques typically generate powerful black-box models that work efficiently in high dimensions but lack in interpretability and accuracy. This can be a problem when we deal with safety-critical systems where failure can cause serious damages and high costs. On the other hand, logics are languages that are easily interpreted by humans and provide verification algorithms that can check in automatic the value of satisfaction of interesting behaviors. However, they often have scalability issues.

In the last years, we are investigating how we can combine these techniques to improve explainability of ML components and scalibility of formal verification, manly working with Signal Temporal Logic (STL). In [Nenzi *et al.*, 2018], we deal a two-label classification problem: given a set of regular and anomalous trajectories, we design a technique to learn the STL formula that better discriminates between the two label data sets, i.e. a formula that is satisfied (as more as possible) by the regular trajectories and not satisfied by the anomalous ones. In the methodology, we exploit genetic programming, a form of EC, to mine the structure of the formula and Bayesian optimization to learn the parameters. In [Mohammadinejad *et al.*, 2021], we consider instead unsupervised learning: starting from a template formulas, we design an agglomerative hierarchical clustering technique to learn the parameter of the formula in such a way that each cluster satisfies a distinct formula. In [Pigozzi *et al.*, 2021], we propose a methodology, to mine STL formulas (both structure and parameters) from unlabeled real-world road traffic data.

### 3.3 Predictive monitoring of CPS

Runtime predictive monitoring (PM) aims at predicting, at runtime, whether a CPS, in a certain state, is about to encounter a safety violations in a nearby future. Effective PM must balance between prediction accuracy, to avoid errors that can jeopardize safety, and computational efficiency, to support fast execution at runtime.

Neural State Classification (NSC) is a recently proposed method that approximates the PM problem so that it become efficient enough for online applications. NSC trains a DNN as an approximate *reachability predictor* that labels an state $x$ as *positive* if an unsafe state is reachable from $x$ within a given time bound, and labels $x$ as *negative* otherwise. NSC predictors are efficient and have very high accuracy, yet are prone to prediction errors that can negatively impact reliability. To overcome this limitation, we present *Neural Predictive Monitoring* (NPM), a technique that complements NSC predictions with estimates of the predictive uncertainty [Bortolussi *et al.*, 2021; Cairoli *et al.*, 2021]. These measures yield principled criteria for the rejection of predictions likely to be incorrect, without knowing the true reachability values. We also present an active learning method that significantly reduces the NSC predictor's error rate and the percentage of rejected predictions. We develop two versions of NPM, a frequentist and a Bayesian approach, based respectively on the use of conformal predictions and Bayesian neural networks as measures of the predictive uncertainty. NPM has been extended to work under more realistic settings, where only partial and noisy observations of the state are available at runtime.

### 3.4 Adversarial robustness of Bayesian Deep Learning

Despite their huge capabilities on a variety of applications, the adoption of Deep Neural Networks (DNNs) in safety-critical applications is significantly limited by their black box nature, i.e., by the lack of explanations underlying predictions. Another major drawback of DNNs is their vulnerability to adversarial attacks, which are small perturbations of test inputs that produce highly confident misclassifications. We address both problems in the context of image classification, with a focus on gradient-based adversarial attacks. Leveraging the manifold hypothesis, we analyze the geometry of adversarial attacks in the large-data, overparameterized limit and prove that, in this limit, Bayesian Neural Networks are robust to gradient-based attacks. Despite the strong theoretical assumptions and the use of approximate Bayesian inference algorithms, we experimentally show that BNNs provide both high accuracy and robustness to gradient-based attacks [Carbone *et al.*, 2020]. Finally, we consider the problem of the stability of saliency-based explanations of DNN predictions under adversarial attacks [Carbone *et al.*, 2021]. In this setting, we show that Bayesian interpretations are considerably more stable compared to those provided by their deterministic counterparts and by adversarially trained networks. Additionally, we provide a theoretical justification in terms of the geometry of the data manifold and also discuss the stability of high level representations of the inputs in the internal layers of the NNs.

## References


[Bartoli *et al.*, 2016] Alberto Bartoli, Andrea De Lorenzo, Eric Medvet, Dennis Morello, e Fabiano Tarlao. " best dinner ever!!!": Automatic generation of restaurant reviews with lstm-rnn. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 721–724. IEEE, 2016.

[Bartoli *et al.*, 2017] Alberto Bartoli, Andrea De Lorenzo, Eric Medvet, e Fabiano Tarlao. Active learning of regular expressions for entity extraction. *IEEE transactions on cybernetics*, 48(3):1067–1080, 2017.

[Bartoli *et al.*, 2018] Alberto Bartoli, Mauro Castelli, e Eric Medvet. Weighted hierarchical grammatical evolution. *IEEE transactions on cybernetics*, 50(2):476–488, 2018.

[Bartoli *et al.*, 2019] Alberto Bartoli, Andrea De Lorenzo, Eric Medvet, e Fabiano Tarlao. Automatic search-and-replace from examples with co-evolutionary genetic programming. *IEEE transactions on cybernetics*, 2019.

[Bortolussi *et al.*, 2021] Luca Bortolussi, Francesca Cairoli, Nicola Paoletti, Scott A. Smolka, e Scott D. Stoller. Neural predictive monitoring and a comparison of frequentist and bayesian approaches. *Int. J. Softw. Tools Technol. Transf.*, 23(4):615–640, 2021.

[Cairoli *et al.*, 2021] Francesca Cairoli, Luca Bortolussi, e Nicola Paoletti. Neural predictive monitoring under partial observability. In Lu Feng e Dana Fisman, editors, *Runtime Verification - 21st International Conference, RV 2021, Virtual Event, October 11-14, 2021, Proceedings*, volume 12974 of *Lecture Notes in Computer Science*, pages 121–141. Springer, 2021.

[Canfora *et al.*, 2015] Gerardo Canfora, Andrea De Lorenzo, Eric Medvet, Francesco Mercaldo, e Corrado Aaron Visaggio. Effectiveness of opcode ngrams for detection of multi family android malware. In *2015 10th International Conference on Availability, Reliability and Security*, pages 333–340. IEEE, 2015.

[Carbone *et al.*, 2020] Ginevra Carbone, Matthew Wicker, Luca Laurenti, Andrea Patane, Luca Bortolussi, e Guido Sanguinetti. Robustness of bayesian neural networks to gradient-based attacks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, e H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15602–15613. Curran Associates, Inc., 2020.

[Carbone *et al.*, 2021] Ginevra Carbone, Guido Sanguinetti, e Luca Bortolussi. Resilience of bayesian layer-wise explanations under adversarial attacks. *CoRR*, abs/2102.11010, 2021.

[Castelli *et al.*, 2021] Mauro Castelli, Luca Manzoni, Tatiane Espindola, Aleš Popovič, e Andrea De Lorenzo. Generative adversarial networks for generating synthetic features for wi-fi signal quality. *Plos one*, 16(11):e0260308, 2021.

[De Lorenzo *et al.*, 2020] Andrea De Lorenzo, Fabio Martinelli, Eric Medvet, Francesco Mercaldo, e Antonella Santone. Visualizing the outcome of dynamic analysis of android malware with vizmal. *Journal of Information Security and Applications*, 50:102423, 2020.

[Manzoni *et al.*, 2020] Luca Manzoni, Alberto Bartoli, Mauro Castelli, Ivo Gonçalves, e Eric Medvet. Specializing context-free grammars with a (1+ 1)-ea. *IEEE Transactions on Evolutionary Computation*, 24(5):960–973, 2020.

[Mohammadinejad *et al.*, 2021] Sara Mohammadinejad, Jyotirmoy V. Deshmukh, e Laura Nenzi. Mining interpretable spatio-temporal logic properties for spatially distributed systems. In Zhe Hou e Vijay Ganesh, editors, *Automated Technology for Verification and Analysis - 19th International Symposium, ATVA 2021, Gold Coast, QLD, Australia, October 18-22, 2021, Proceedings*, volume 12971 of *Lecture Notes in Computer Science*, pages 91–107. Springer, 2021.

[Nenzi *et al.*, 2018] Laura Nenzi, Simone Silvetti, Ezio Bartocci, e Luca Bortolussi. A robust genetic algorithm for learning temporal specifications from data. In *International Conference on Quantitative Evaluation of Systems*, pages 323–338. Springer, 2018.

[Pigozzi *et al.*, 2021] Federico Pigozzi, Eric Medvet, e Laura Nenzi. Mining road traffic rules with signal temporal logic and grammar-based genetic programming. *Applied Sciences*, 11(22):10573, 2021.

[Salvato *et al.*, 2021] Erica Salvato, Gianfranco Fenu, Eric Medvet, e Felice Andrea Pellegrino. Crossing the reality gap: a survey on sim-to-real transferability of robot controllers in reinforcement learning. *IEEE Access*, 2021.

[Talamini *et al.*, 2020] Jacopo Talamini, Alberto Bartoli, Andrea De Lorenzo, e Eric Medvet. On the impact of the rules on autonomous drive learning. *Applied Sciences*, 10(7):2394, 2020.

[Talamini *et al.*, 2021] Jacopo Talamini, Eric Medvet, e Stefano Nichele. Criticality-driven evolution of adaptable morphologies of voxel-based soft-robots. *Frontiers in Robotics and AI*, 8:172, 2021.

[Virgolin *et al.*, 2020] Marco Virgolin, Andrea De Lorenzo, Eric Medvet, e Francesca Randone. Learning a formula of interpretability to learn interpretable formulas. In *International Conference on Parallel Problem Solving from Nature*, pages 79–93. Springer, 2020.

[Virgolin *et al.*, 2021] Marco Virgolin, Andrea De Lorenzo, Francesca Randone, Eric Medvet, e Mattias Wahde. Model learning with personalized interpretability estimation (ml-pie). page 1355–1364, 2021.