# COMPUTER AND CONTROL ENGINEERING

## Responsible AI: from principles to practices

| | |
|---|---|
| **Funded By** | Dipartimento DAUIN<br>FONDAZIONE CRT CASSA DI RISPARMIO DI TORINO<br>[P.iva/CF:06655250014] |
| **Supervisor** | DE MARTIN JUAN CARLOS - juancarlos.demartin@polito.it |
| **Contact** | VETRO' ANTONIO - antonio.vetro@polito.it |
| **Context of the research activity** | The Ph.D. proposal focuses on the problem of accountability of Artificial Intelligence (AI) applications which automate decisions on relevant aspects of human lives. These kinds of automated decisions often unfairly discriminate against certain individuals or groups of individuals, on grounds that are unreasonable or inappropriate. The goal of the research is to translate the principles and guidelines of responsible and human-centric AI into techniques and actionable industrial practices. |
| | The proposal has two main objectives, namely O1 and O2.<br><br>O1) Research, development, and test of techniques for bias and data quality improvement.<br><br>The Ph.D. candidate will research, implement, and test techniques that can help detect bias and data quality issues in training data and mitigate them by experimenting with a variety of techniques.<br>To reach this goal, the Ph.D. candidate will research and experiment on the most suitable measures of data bias and data quality and on intervention mechanisms to improve them, with a special focus on pre-processing techniques (e.g., data rebalancing algorithms, datasets augmentation methods). Quality of metadata and of dataset documentation will be also thoroughly investigated.<br>The high-level research work plan for O1 is described herein.<br><br>O1.A Datasets:<br>a) Identification of data quality and bias measures (created ad-hoc and searched in the literature, also from other disciplines), and tests on available datasets, on their mutations and on synthetic datasets.<br>b) Experiments on the propagation of bias and quality issues to the output of classification/prediction tasks, with different algorithms and datasets.<br>c) Correlational analyses of bias and quality measures with classification output measurers, with special attention on the fairness-accuracy tradeoff.<br>d) Identification and test of mitigation techniques, through re-iteration of steps b) and c) |

| | |
|---|---|
| **Objectives** | O1.B Datasets documentation:<br>a) Identification of guidelines and measures for dataset documentation quality<br>b) Setup and execution of measurements on available datasets<br>c) Analysis of results and of possible consequences as “data cascades”<br><br>O1.C. Data labeling: based on the results of previous steps, design and prototype of informative ethical-sensitive data labels that can inform stakeholders (data maintainers, model builders, end users, etc.) about the risk of downstream effects from early data issues in the AI pipelines. The data labeling schemes will be designed and tested with the goal to facilitate early interventions and mitigations of data cascades, encompassing both human intervention (through interactive visualizations) and seamless introduction in the AI pipeline.<br><br>O2) Research, development, and test of model enrichment methods to facilitate human comprehension, intervention and overall agency in AI model development and monitoring.<br><br>While generating explanation-based assessments for black-box AI models is not very complicated, presenting that information to humans in a way that they can make positive interventions within the datasets and models is an open research issue. In addition, most of the available approaches to implement eXplainable Artificial Intelligence (XAI) focus on technical solutions usable only by experts able to manipulate the algorithms, but not by other relevant stakeholders (including end users). The Ph.D. candidate will research, implement, and test the best practices to provide augmented cues not only to model builders but also to laymen. In this context, a promising alternative to the most beaten routes for XAI is represented by Knowledge Graphs, because they are natively developed to support explanations intelligible to humans: therefore, the Ph.D. candidate will research on how to integrate symbolic systems in the typical AI pipeline for the purpose of facilitating human comprehension, intervention, and overall agency.<br><br>The high-level research work plan for O2 is described herein.<br><br>a) Research, development (and/or reuse and adaptation) and testing of a knowledge matching mechanism to bind input features of selected classification/prediction algorithms to classes of an ontology and entities of a knowledge graph (KG). The process could be partially or fully automated. The activity may include the building of an ontology and of a KG in case they are not yet available for the scenario being tested.<br><br>b) Research, development (and/or reuse and adaptation) and testing of an interactive application for producing explanations for algorithms output, by leveraging the manipulation of symbols from the previously developed KG, to produce i) transparent inferences of new information and ii) high level explanations. The interaction with the user should be in the form of questions made by users in natural language (also using predefined questions) that are translated in SPARQL queries run over the knowledge graph. The visual output should be evaluated in terms of usability, level of comprehension and possibility to configure changes in the model and/or in the data.<br><br>The Ph.D. work will be done in strict collaboration with the company ClearboxAI, startup incubated in I3P, winner of the National Innovation Award |

| | |
|---|---|
| | (PNI 2019) in the ICT category and of the EU Seal of Excellence. ClearboxAI will share know-how and access to its technological assets with the Ph.D. candidate, to make research experiments within real industrial settings and scenarios. |

| | |
|---|---|
| **Skills and competencies for the development of the activity** | Engineering competences: basic knowledge of working mechanisms of machine learning and artificial intelligence techniques, supervised/unsupervised/deep learning; techniques for data analysis and visualization.<br><br>Other competences: predisposition into interdisciplinary research for understanding ethical issues in digital technologies. Capability to understand law texts. |