**COMPUTER AND CONTROL ENGINEERING**

**Explainable AI for spoken language understanding**

**Funded By**

Dipartimento DAUIN  
Centro Interdipartimentale SmartData@PolIITO

**Supervisor**

BARALIS ELENA MARIA - elena.baralis@polito.it

---

**Context of the research activity**

Machine learning models and automated-decision making procedures are becoming more and more pervasive, thus a growing interest is arising on the careful understanding of their behavior. The focus of this research activity is on models for spoken language understanding.

Spoken language understanding (SLU) systems infer semantic information from user utterances. Traditionally, SLU systems rely on two models: the automatic speech recognition and the Natural Language Understanding models. The former processes speech signals and translates the utterance to text, while the latter processes the text to derive the target internal representation. This pipeline allows to clearly separate the two processes and validate the corresponding models separately. In contrast, end-to-end (E2E) models directly rely on spoken signals to infer the utterance semantics.

E2E SLU models impose new challenges for the explainable AI research. These models do not reveal the reasons behind predictions and their results are hence hard to interpret. Moreover, the combination of both the acoustic and semantic information into a single model makes the identification and the understanding of errors more difficult.

---

**Objectives**

End to End (E2E) Spoken Language Understanding (SLU) models perform the spoken language understanding task as a complex black-box process without the need of distinct automatic speech recognition and natural language understanding steps. The semantics of the utterance are inferred without showing any intermediate step, as for example explicit transcription of strings. Hence, explaining model errors and understanding the reasons for its performance becomes a difficult task.

Investigating the presence of data subgroups that behave in problematic ways is central to model understanding, as well as to studying model fairness, and debugging AI pipelines. Typically, the overall performance of an AI model reveals how well the model performs on average on the whole dataset. However, it does not reveal the problems that may affect particular portions of the data.

The research activity will address the explanation of an E2E SLU model by identifying and characterizing data subgroups for which model performance shows an anomalous behavior (e.g., the False Positive Rate is higher than
average). Critical subgroups may be identified by exploiting the notion of pattern. Patterns are conjunctions of attribute-value pairs intrinsically interpretable.

The research activity will consider model agnostic techniques, because they do not rely on the knowledge of the inner workings of any classification paradigm. Since pattern mining techniques do not rely on any specific language knowledge, they may provide an effective tool to address different spoken languages.

<table>
<thead>
<tr>
<th>Skills and competencies for the development of the activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine learning, data science and data analysis</td>
</tr>
</tbody>
</table>