ASPETTANDO LA
NOTTE EUROPEA DEI RICERCATORI

Symbolic Knowledge Injection and Extraction for Autonomous Learninig

Matteo Magnini

Department of Computer Science and Engineering, PhD in Computer Science and Engineering 38° cycle, supervisor Prof. Andrea Omicini

Background

Artificial intelligence (AI) has evolved from early symbolic reasoning systems to modern data-driven neural networks, enabling automation of cognitive tasks but raising challenges of interpretability, reliability, and alignment. **Neuro-symbolic AI (NeSy)** addresses these issues by integrating the deductive reasoning of symbolic methods with the inductive learning capabilities of sub-symbolic models. Within this paradigm, **symbolic knowledge injection (SKI)** and **symbolic knowledge extraction (SKE)** are two complementary pillars: SKI injects prior knowledge into models to guide learning and improve robustness, while SKE extracts interpretable, reusable knowledge from tra ined systems. Recent advances in large language models (LLMs) create new opportunities for hybrid AI, fostering systems capable of autonomous reasoning and learning in complex domains.

Project Goals

- **Advance SKI and SKE techniques**: Provide a systematic taxonomy, propose new methodologies, and develop evaluation metrics;
- **Develop engineering tools**: Build software libraries and frameworks (e.g., PSyKI) to facilitate SKI/SKE integration into real-world AI pipelines;
- **Enable transparent and robust AI**: Design methods to improve trustworthiness and fairness in machine learning models, with applications in healthcare and intelligent multi-agent systems;
- **Support autonomous learning**: Lay foundations for AI systems that can learn, reason, and adapt with minimal supervision, contributing toward Artificial General Intelligence (AGI).

Experimental Approach

We decompose the challenge of building autonomous AI systems into modular research tasks:

1. **Formal analysis**: Systematic review and categorization of SKI and SKE approaches, considering knowledge representation, integration strategies, and scalability;
2. **Algorithm design**: Novel injection methods (e.g., KILL, KINS) and extraction pipelines for interpretable logic and ontological knowledge;
3. **Platform development**: Creation of PSyKI, an open-source platform enabling rapid prototyping of neuro-symbolic workflows;
4. **Domain studies**: Application of methods in explainable recommendation systems, fairness-aware ML, and medical chatbots powered by Retrieval-Augmented Generation (RAG);
5. **Evaluation**: Experiments on robustness, constraint enforcement, knowledge fidelity, and performance trade-offs across synthetic and real-world datasets.

Expected Outcomes

- A **comprehensive taxonomy** of SKI and SKE techniques to guide research and practice.
- Novel **knowledge injection mechanisms** with demonstrated performance improvements and robustness under noisy conditions;
- **PSyKI platform**: A software library providing unified APIs and workflows for symbolic-neural integration;
- Proven strategies for **fairness-by-design** through constraints injection, and interpretable pipelines for knowledge extraction;
- Demonstrated **real-world impact:** nutritional recommenders, ontology population and learning with LLMs, medical RAG applications;
- Contributions published in peer-reviewed venues, reinforcing the role of SKI and SKE in the future of explainable and autonomous AI.

SUSTAINABLE DEVELOPMENT GOALS