Abstract for the general public

The project concerns **Machine Learning**, a branch of **Artificial Intelligence** (AI) that deals with algorithms that autonomously accumulate knowledge from various data sources. The main driving force behind contemporary Machine Learning is **Deep Learning** (DL): its recent progress caused many AI applications to reach their 'tipping points' and turned them from mere proofs-of-concept into useful products and services that now benefit billions of users worldwide.

However, DL also suffers from **fundamental limitations**, among others (i) insatiable appetite for data and computing power, which leads to exorbitant energy consumption and inclines the key players in the AI market consider nuclear reactors for powering their servers, and (ii) limited capability of coherent, transparent, interpretable and reliable reasoning, which leads to errors, biases, and hallucinations, the latter becoming recently particularly evident in Large Language Models (LLMs). These downsides have **severe implications in real-life applications**, like faulty or biased decision-making and flawed medical diagnosing. Therefore, the challenges addressed in this project are prospectively relevant not only for the scientific community but also for regular users of AI-based solutions, viz. for everyone.

To address these challenges, this project resorts to **neurosymbolic systems** which are hybrids of conventional DL with symbolic, more explicit processing of information that takes place on a higher abstraction level. Such systems combine the best of both worlds, i.e. the known advantages of DL with the crisp, transparent, and versatile symbolic inference. In contrast to purely DL-based solutions, neurosymbolic systems are much more transparent, interpretable, and require less data and computation for training and usage. For instance, within preparations for this project, the project team has proposed a specific class of neurosymbolic systems which, even when trained on relatively few examples of visual scenes, was capable to interpret them robustly and provide similar capacity for new scenes, i.e. make correct predictions when confronted with previously unseen data, viz. generalize. They demonstrated these advantages on a challenging problem of medical diagnosing based on microscopic imaging of human thyroid (histological sections), building more accurate predictive diagnostic models than the conventional DL.

The project will result in new AI algorithms that are more efficient, require less data for training and explain their actions better, and so contribute to both the **theory and practice of AI**. Concerning the former, it will expand the foundations of AI and provide the community of researchers and practitioners with novel algorithms and new ways in which they interact with humans – in particular, how they explain their reasoning. Concerning the latter, it will validate the proposed methods on real-world problems in, among others, (i) **biomedicine**, which notoriously suffers from insufficient data and high costs of their annotation while requiring explanations of decisions made for individual patients and of entire algorithms (e.g. for regulatory compliance), and (ii) **satellite imaging of the Earth**, where quick progress is essential to improve our understanding of the dynamic, rapid changes that currently affect our planet. These proof-of-concept demonstrators only scratch the surface of the spectrum of potential practical implications of the project, which may reach to many other areas like environmental science, engineering of renewable energy sources, and more.

The results of the project will be disseminated via publications in highly-ranked scientific journals, presentations at conferences and other scientific events, a continuously maintained webpage dedicated to the project, and social media. On top of that, the main tangible outcome of the project will be **a software library available under an open-source license to the global scientific community and beyond**.