Abstract for the general public

Recent advancements in AI resulted in a tremendous number of methods that can be trained to mimic human behaviour in the numerous tasks. We already have autonomous vehicles, machines that can replace human operators in simple tasks such as checking the quality of data samples or produced goods, or even more complex systems where AI powered computers can automatically calculate insurance rates more precisely than any expert. The underlying assumption behind the success of such systems is that we can train them on top of vast amounts of carefully labelled examples. While such a situation greatly simplifies the problem to the issue of learning to recognise specific patterns, it does not necessarily lead to universal methods. That so called *supervised learning* paradigm is a bottleneck for building more intelligent generalist models. Practically speaking, it's impossible to label everything in the world.

In our early childhood, we not only learn through guidance from our supervisors. We explore the world around us in an unsupervised manner and learn by experiencing our surroundings on our own (often based on our own mistakes). Machine learning concepts that try to mimic such a way of learning are grouped around the term of *self-supervised learning*.

However, contrary to children, all artificial neural structures suffer from *catastrophic forgetting* - a situation where the performance of a model drops significantly every time it is retrained with new information. For instance, if a network previously trained for detecting virus infections is now retrained with data describing a recently discovered strain, the diagnostic precision for all previous ones drops significantly. In human terms, it is as if we forget every single experience whenever we adjust to the new task we have to perform.

In this project, we postulate that self-supervised learning is a crucial area of Artificial Intelligence that might be a key ingredient in discovering systems that are less prone to forgetting. We already know that knowledge discovery in self-supervised systems is more similar to the way we learn as humans. Hence, on top of this observation, our previous works and preliminary results, we hypothesise that this difference might also be observed in the way machine learning systems forget. Therefore, we plan to at first better understand the nature of forgetting in self-supervised learning systems in order to propose novel solutions that will continually accumulate knowledge embedded in the real world around us without supervision.