Among various approaches to *artificial intelligence* (AI), *machine learning* (ML) is particularly effective. What distinguishes some of its algorithms is the *black box problem*, which arises with the use of epistemically opaque functions. These functions are often too complex for us to analyze, and it may not be possible to fully understand AI's decision-making criteria. To restore trust in automated decision making, researchers are currently focusing on the *explainable AI* (XAI) research program, which seeks explainability, transparency, and interpretability of algorithms as ways of validating the decision-making processes of epistemically opaque AI systems.

While there has been a lot of technical work on the subject already, it often lacks solid conceptual foundations and is not well integrated with research on *understanding* or the notion of scientific *explanation* itself. In this light, the aim of the project is to philosophically analyze the black box problem in machine learning and to explicate the concept of AI explainability. The realization of this goal will be accompanied by an examination of the relationship of the relationship between explanation and understanding in epistemically opaque ML models, drawing from the models of explanation in the philosophy of science and epistemological theories of understanding.

The main question posed in the project is the question of how explanations enable users to understand the operation of opaque AI systems. As we do not have precise criteria for the adequacy of an AI explanation, the goal is also to elucidate a set of theoretical requirements for a theoretically informed notion of AI explainability that could play a role in the evaluation of XAI models. Solving these problems will involve finding answers to several specific research questions, which include the following: What are the criteria for AI explainability? What theoretical and practical requirements should be met in a "good" explanation? What kind of understanding of AI models do and should XAI solutions deliver? What kind of knowledge do users care about gaining in terms of interpreting how AI works?

In order to conduct empirical research, scientists need concepts, without which it is impossible to talk about knowledge and communicate it to those who want to learn about the reality around us. Consequently, the primary research method used in the project is conceptual analysis, focusing on the identification and clarification of concepts in terms of their clarity, precision, coherence or empirical relevance, as well as conceptual engineering, which aims to correct the meanings of existing concepts and identify the meanings of terms we should use to better illuminate and solve the conceptual and theoretical problems we encounter in scientific practice. In studying this practice, existing technical methods for producing explanations in XAI will be critically evaluated and contrasted with various kinds of explanations considered in the philosophy of science. Through the clarification of concepts, revealing implicit assumptions, and scrutinizing the practices of researchers, philosophy of science can make an important contribution to scientific debates.

The project will examine, among other things, *mechanistic explanations* developed within the New Mechanism tradition, which are descriptions of interactions of operating parts and the internal organization of mechanisms responsible for phenomena; *topological explanations* postulated by Daniel Kostic, which describe how the mathematical properties of connection patterns in complex networks determine the dynamics of systems exhibiting those patterns; or *counterfactual explanations*, which aim to reveal variables that should be different in the initial conditions of a phenomenon in order to observe an expected outcome.

The purpose of explaining an AI model is to make its behavior understandable to its users. But without a previous grasp of what it means to say that an agent understands a model or decision, the explanatory strategies will lack a well-defined goal. This is why the project also seeks to develop a conceptual and theoretical apparatus for the issue of understanding AI explanations. In recent years, the relationship between explanation and understanding has been studied in detail by Henk de Regt, Kareem Khalifa and Daniel Wilkenfeld, whose approaches could be used to integrate a range of explanatory strategies in terms of their potential to produce authentic understanding in users. A potentially conciliatory approach to understanding AI explanations while taking into account various epistemic perspectives of AI users can be found in Michela Massimi's *perspective realism*.

The results of the project will help to problematize the relationship between explanation and understanding in XAI, offering a novel philosophical diagnosis of the problem. Besides theoretical insights, it will provide a typology of AI explanations offered by a range of methods consistent with the literature. Along with the rapid and widespread development of intelligent technology, we are experiencing the growing ethical risk of high-stake opaque decision making, which may eventually lead to socially unacceptable, or even catastrophic outcomes. The gap in the research on AI explainability needs to be addressed with a high priority to ensure the development of responsible and beneficial technology for the future. This task requires multidisciplinary focus with expertise not only in computer science or other hard sciences, but also in philosophy and humanities, in order to lay a solid theoretical groundwork foundation for further technological progress in the field of artificial intelligence.