Käte Hamburger Kolleg: Cultures of Research
Lecture Series 2022 “Philosophy of AI_Optimits and Pessimist Views”

Lena Kästner (University of Bayreuth), 15. June 2022

Explaining AI Through the Scientific Perspective

Modern artificial intelligence (AI) systems are often complex and opaque. At the same time, they are becoming increasingly prevalent in our lives. As a result, there is an increasing demand to make AI systems *explainable*and their behaviour *intelligible*. Recent work primarily approaches this problem by employing specific explainability methods to aid in-context understanding. We think, however, that important desiderata such as safety and reliability might be best satisfied through the increase of expert understanding with respect to *how AI systems work*.

We suggest this requires research from the scientific perspective. This approach starts from the premise that once AI systems become sufficiently complex, they are best investigated and explained through the same lens as biological organisms. Accordingly, this work seeks to characterise the functional structure that emerges in AI systems through training. As we will describe, researchers pursuing this approach adopt strategies for discovery that have proved successful in the life sciences, such as pattern recognition, functional decomposition, localization, and systematic manipulation.

In this talk, we discuss the promises and limitations of the scientific approach to understanding AI systems. We contend that uncovering their emergent structure plays an important and underappreciated role in solving the explainability problem in AI.