**TITLE:** ESTABLISHING RELIABILITY WHILE EMBRACING UNEXPLAINABILITY: *Moving beyond explainability as a criterion for ML models to produce reliable and justified knowledge.*

**ABSTRACT:**

Central to scientific knowledge production is not just the empirical adequacy of knowledge claims but also the reliability and justification of the tools and methodologies employed in the production of knowledge claims. In the context of contemporary machine learning research, establishing the reliability and justification of knowledge claims produced using ML models has remained a major challenge given that ML researchers, despite having achieved significant breakthroughs in various  computational tasks, were often unable to identify and explain which factors are responsible for empirical gains, a problem stemming from the opaqueness of ML models and the complexity of the tasks they are applied to. The failure to identify sources for empirical gains and increasing presence of ill-understood techniques (often referred to as the model zoo) have raised concerns.

Emerging developments in Explainable AI can be seen as ways of confronting these theoretical shortcomings of ML models. XAI aims to render the modes of operation of ML models transparent so that these modes can later be assessed by particular scientific disciplines in accordance with their theoretical and methodological resources that are specific to them. Do ML Models necessarily have to be rendered explainable for them to be incorporated in scientific practices and their knowledge claims to be considered reliable? Having explainability as a rigid requirement for scientific use would render a majority of state-of-the-art models (which have portrayed significant empirical gains in various computational tasks) inaccessible to scientists. Can we identify ways in which black-boxed models can still be used as reliable instruments in science while maintaining their opaque nature? This paper aims to address these questions by providing an alternative approach to establishing the reliability of ML driven knowledge production, one that isn’t confined by the opaque nature of ML models.

In order to make my case, I first draw on instances from history of science where technologies have contributed to knowledge production despite there being no established theoretical justification (e.g., the case of the development of the electric motor of Thomas Davenport which preceded the development of electromagnetic theory). Secondly, I illustrate how opaqueness remains a valid concern only in cases where scientists are considering the predictions of the model themselves as candidates of scientific knowledge claims that stand in need for justification. However, as I will show in my essay, there are alternate spaces within scientific practices whose ends are not necessarily the production of knowledge claims themselves but those that nevertheless play an instrumental role in knowledge production. ML models, I argue, can be instrumental in such spaces in the production of reliable knowledge despite them being opaque and despite there being no bridging between the theoretical resources of specific scientific disciplines and the ML model.