Abstract for the Thesis Submitted in Fulfillment of the Requirements for the

Degree of Doctor in the Interdisciplinary Studies: Business Economics, Science and Sociology

Black Box Revelation: Interdisciplinary Perspectives on Bias in AI

Carmen Mazijn

Abstract

The integration of AI systems in decision-making processes often promises increased efficiency, accuracy, and objectivity through automation and standardisation. However, the deployment of these systems has raised critical concerns regarding fairness, transparency, and accountability. Biased algorithms have the potential to cause allocative and representational harms through discrimination and stereotyping, consequently exacerbating social inequalities.

Despite the presence of anti-discrimination laws, the detection of algorithmic bias is challenging due to the inherent opaqueness of AI systems, limited access to datasets and models, the use of proxies to encode protected characteristics, and the prevalence of intersectional discrimination. In this context, examining the interplay between AI technologies and decision-making processes is crucial.

In this dissertation, we address immediate and long-term harms caused by AI systems interacting with our world. We argue for a paradigm shift towards equality of treatment, particularly through the development of input-based group fairness evaluation methods. Furthermore, we provide a taxonomy of AI-driven feedback loops and their potential long-term consequences. Finally, to ensure better utilization of these tools, we have composed policy recommendations and introduce a design thinking approach for developers.