LEIBNIZ UNIVERSITÄT HANNOVER
WIRTSCHAFTSWISSENSCHAFTLICHE FAKULTÄT
INSTITUT FÜR WIRTSCHAFTSINFORMATIK

-

How to Make AI-based Decisions More Explainable - An XAI Implementation Process Model

-

Masterarbeit

zur Erlangung des akademischen Grades „Master of Science (M. Sc.)“ im Studiengang Wirtschaftsingenieur
der Fakultät für Elektrotechnik und Informatik, Fakultät für Maschinenbau und der Wirtschaftswissenschaften Fakultät der Leibniz Universität Hannover

vorgelegt von

Name: Hojeij Vorname: Rawad

Geb. am: in:

Prüfer: Prof. Dr. M. H. Breitner

Ort, Datum: Hannover, den 17.05.2023

Table of Contents

Research Summary

1. Introduction

"With artificial intelligence we are summoning the demon."
-Elon Musk

The artificial intelligence (AI) arms race has understandably attracted media attention, capturing the interest of many. It is common for revolutionary technologies to elicit predictions of catastrophe, and AI is no exception. Influential figures like physicist Stephen Hawking and entrepreneur Elon Musk, the founder of SpaceX and Tesla, have expressed concerns regarding the existential risks posed by highly intelligent AI systems (Gent, 2015, p. 2). The use of AI as a decision assistant in critical domains such as medicine, law and autonomous driving systems has become widespread (Jin, et al., 2023, p. 1). As the dependence on decision-making systems increases, so does the complexity of these systems, the majority of which rely on machine learning (ML) and exhibit a "black box" nature (Schaaf, et al., 2021, p. 11). This has led to the development of explainable artificial intelligence (XAI) systems that aim to provide clear explanations for their recommendations and decision-making processes (Chazette, et al., 2022, p. 1). Although, the wide range of different methods developed for black-box models and the lack of a practical standard implementation complicate the selection of the most appropriate approach for a particular application (Schaaf, et al., 2021, p. 12). Additionally, explainability depends on how the recipient perceives and understands the explanation, as well as other human factors. The complete transparency of the model alone is no guarantee that the receiver can fully understand all the information (Liao & Varshney, 2021, p. 2).
Therefore, for the successful development of XAI-Software, it is essential to adopt a human-centered approach that prioritizes the explainability requirements of the users while simultaneously focusing on the technical requirements and capabilities. The integration of both approaches within the software development life cycle (SDLC) remains an area that is largely unexplored. To close this research gap, we pose the following research questions (RQs):

> RQ 1: What are the key steps required to implement XAI?
>
> RQ 2: How can these steps be integrated into the SDLC for the development of XAI systems?

The objective of the RQ1 is to provide a broad perspective on successful activities and techniques that facilitate the implementation of explainable systems. Building on the

8 Conclusion

The increasing emphasis on explainability in AI research is due to its critical role in addressing ethical concerns and promoting trust in the real-world application of AI systems. To improve the explainability of AI-based decisions and promote the successful implementation of XAI systems, we conducted an extensive literature review. By applying Hevner's (2007) DSR methodology, we have gained valuable insights that guided the development of our design artifact, the XAI-PM, which aligns with the SDLC. Building upon the knowledge gleaned from literature review, we identified ten essential steps and assessed their practicality and feasibility through an interview study with five experts from the fields of software development and AI.

In light of the findings, the design artifact highlights important outcomes. This work strives to facilitate the incorporation of XAI into practical applications by offering essential insights into the necessary steps within the SDLC to achieve a successful implementation of XAI. The contribution to theory and practice extends the current knowledge base by emphasizing the significance of considering both user-centered approach and the technical requirements during the selection and implementation of the XAI method. In pursuit of this goal, we have proposed thirty-one XAI methods and examined their alignment with the previously gathered requirements. Additionally, we emphasize the significance of co-design and prototyping steps in ensuring user satisfaction. In the implementation phase, we also examined which XAI methods could be effectively implemented using specific libraries. Furthermore, our evaluation phase provides an opportunity to assess both the system's interaction with the user and the quality of the explanations it provides. However, due to the maturity of our artifact, numerous research avenues emerge for the different phases of the SDLC, as depicted in Table 11.

Moreover, due to the limited scope of evaluation conducted on this artifact, the definitive quality and usefulness of the design artifact could not be ascertained. While XAI has gained traction in various application domains, the literature lacks sufficient real-world implementation experiences and coverage across all phases of the SDLC. Consequently, this master thesis suggests the need for a more comprehensive evaluation of the results through a case study. In this regard, we view our proposal as a critical first step and a foundation that can be refined and extended in future research efforts.