Leibniz Universität Hannover
Wirtschaftswissenschaftliche Fakultät
Institut für Wirtschaftsinformatik

Artificial Intelligence: A Use Case Based Analysis on Trustworthiness, Verification, and Certification

Masterarbeit

zur Erlangung des akademischen Grades „Master of Science (M.Sc.)“ im Studiengang Wirtschaftswissenschaft der Wirtschaftswissenschaftlichen Fakultät der Leibniz Universität Hannover

vorgelegt von:

Name: Nowikow [redacted] Vorname: Leon [redacted]

Prüfer: Prof. Dr. M. H. Breitner

Hannover, den 01.04.2025

Contents

Research Summary

Introduction

In the age of rapid technological progress, artificial intelligence (AI) is becoming increasingly important (Maslej et al., 2024). Among these technologies, Large Language Models (LLMs) such as GPT-4, Claude or Gemini stand out in particular, offering a wide range of possible uses through their ability to process and generate language (Maslej et al., 2024). These models are increasingly being used in many fields, from automated customer communication to supporting scientific research (Enholm et al., 2022; Jason Furman and Robert C. Seamans, 2018). However, the widespread deployment of LLMs also raises critical questions about their Trustworthiness.

In particular, generative AI (GenAI), which are partly based on LLMs, offer a wide range of possible applications such as the generation of text, code, music or images (Maslej et al., 2024). Due to the broad database on which such a model is trained, the possible fields of application are difficult to narrow down. They can be used for entertainment purposes, as knowledge databases, coding aids or even to provide support in safety-critical areas such as medicine, law or finance. The broad user base of such systems leads to specific challenges in several areas related to technical and ethical requirements, since these technologies are used by both experts and inexperienced user groups. The easy accessibility of GenAI tools increases the risk of misuse, unintentional security breaches, and the spread of erroneous or manipulative content. Furthermore, the large database is also a cause for concern. Due to the use of billions of parameters within a LLM (D. Kim et al., 2023; Touvron et al., 2023), there are concerns regarding the assurance of data quality and the resulting outputs of a system.

Trustworthiness, as a concept for evaluating the reliability, security, fairness and transparency of AI systems, is a central subject of research and initial regulatory approaches. Particularly in the context of LLMs and GenAI systems, these criteria are difficult to grasp and often insufficiently defined (Maslej et al., 2024). Therefore, an essential part of this work is to further evaluate the concept of Trustworthiness and to clearly identify central aspects.

Based on this, the extent to which existing certification and verification approaches for AI systems can be applied to LLMs will be investigated, or whether significant gaps exist that require new approaches. Since current regulations and standards are usually not sufficiently designed for these technologies (Maslej et al., 2024), it is also analyzed which adjustments would be useful to enable reliable verification of individual components for companies and independent testing authorities.

Finally, in view of the various use cases and stakeholders, it must be considered how such quality assurance can be clearly communicated to the different user groups. An initial approach to this would be certification. Therefore, the third point of consideration in this work is to what extent certification of AI systems is meaningful and feasible, given the wide range of applications.

We will provide a meaningful recommendation for the topics mentioned by focusing on one use case in the context of the feasibility and meaningfulness of possible implications. With GPT-4 and comparable LLM models as a use case, we see a practical way to focus on the current discussion priorities. Due to the wide range of possible uses of such applications, we also see the chance that recommendations based on this example can be applied to other use cases.

Therefore, our research questions for this work are:
RQ1. What are the challenges and problems in implementing, verifying, and certifying TAI as a concept?
RQ2. What approaches are there to solve these challenges and problems in the context of LLM as a considered use case?

To answer these questions, we choose a qualitative approach to analyze current literature on open questions regarding the classification of Trustworthiness, possible approaches for the verification of Trustworthy Artificial Intelligence (TAI) and considerations regarding possible certification approaches. Since no global standard can be recognized in the regulation of AI so far, we try to sharpen the picture in these categories based on real legislation and guidelines from white papers published by governments and institutions worldwide. Finally, we interviewed seven experts of different fields with experience of using AI to obtain domain-specific assessments and challenges on the topics mentioned, as well as to include possible new approaches.

In the following sections we define key terms, describe the research design, and present the synthesized findings to answer the research questions and contribute practical recommendations.

Theoretical Background

Artificial Intelligence (AI) is generally defined „as the ability of a system to correctly interpret external data, learn from that data, and use that knowledge to achieve specific goals and tasks through flexible adaptation“(A. Kaplan & Haenlein, 2019).
Machine Learning (ML) involves creating algorithms that enable systems to make predictions or decisions based on data, and learn from them. The processes are highly

6 Conclusions

In this work, we analyzed the current state of TAI. Our objective was to identify key problems and challenges in the definition of TAI as a concept, the Verification of TAI components, and the Certification of product (components). To do this, we selected a representative use case in the form of LLM, as it is currently the subject of worldwide discussions around the safe integration of AI into society. This is particularly due to the fact that LLM can process new volumes of data and also have access to a larger knowledge base. At the same time, the complexity of such systems makes it difficult to see the actual processes and procedures within the algorithms. We have consulted relevant literature on the topics of Trustworthiness, Verification and Certification for our approach. Subsequently, we provided an overview of some of the global developments in the regulation of AI products, with a particular focus on the differences in the definition of the TAI concept. Finally, we included the assessments from practice by means of expert interviews and also had existing approaches from the literature and regulatory affairs critically classified with them.

We brought together all the results and obtained a comprehensive overview of the topic. In doing so, we realized that TAI as a concept is still not fully developed and that there are some misunderstandings and terminological difficulties among the general public. From this, we concluded that the priorities for this concept are handled differently worldwide and that there is a general lack of clarity about what requirements can and should be set, as well as what they actually mean when looking at an AI system. We have recognized that there are many different approaches in science for securing measurable metrics for the Verification of AI. However, these are often only applicable in highly abstracted environments and cannot be applied to the scale of AI systems that are already in use. There is also a debate as to whether the use of qualitative means makes sense, or whether the lack of objectivity in some of these methods does not guarantee the quality of a product. In the context of regulation, we have observed many requirements, some of which are already legally binding. The definition of the requirements is often too vague and does not allow a technical determination of whether the criteria are actually met, as long as they have been sufficiently declared at all.

In the context of Certification, we have found that this is a useful way to communicate and ensure the quality standards of TAI to the outside world. However, based on our findings, we appeal to Certification by independent third parties to avoid potential problems such as ethics washing. Literature, regulators and experts largely agree on the classification of AI systems into risk categories in order to set priorities for regulation.

In view of our use case, we were able to see that LLMs fall under almost all requirements for a general TAI, but that Verification of compliance with these criteria is difficult. This is due, on the one hand, to the different data formats and the unclear nature and interpretability of the data bases, and on the other hand, to the scale on which the data volumes move.

We see opportunities in the Certification of frequently used LLMs as foundation models, in order to be able to certify many AI systems that build on them in their foundation to a conscientious extent. This requires clear statements about which LLMs these are and how they can be tested for TAI compliance.

We recognize the complexity of the topic and the problem that, on the one hand, detailed elaborations are necessary for all components of TAI, but on the other hand, only comprehensive analyses can reveal possible interactions.

We appeal to the research community to further elaborate existing approaches and thus to continuously work on the interdisciplinary design of a secure AI landscape for the general public.