

Enhancing Trust in Artificial Intelligence

Prepared by: Lama Mansour Alsubaie

Chapter 1: Introduction

1.1 Introduction

Artificial Intelligence (AI) is transforming how decisions are made in fields such as healthcare, law, and transportation. While these advancements bring efficiency and innovation, the complexity and opacity of AI systems raise concerns about transparency and trust. This chapter introduces the need for building trustworthy AI and sets the stage for addressing the challenges of explainability and ethical alignment.

1.2 Background

The evolution of AI from rule-based systems to deep learning has significantly increased its capabilities. However, as AI becomes more integrated into everyday life, questions about its ethical use, interpretability, and social impact have emerged. The 'black box' nature of modern AI models complicates oversight, particularly in domains with high societal impact.

1.3 Problem Statement

Despite the growing prevalence of AI technologies, there remains a persistent lack of trust among users, policymakers, and even developers, largely due to the opaque nature of many AI systems. In high-stakes environments like medicine, finance, and autonomous systems, this opacity—often referred to as the 'black box' problem—prevents stakeholders from fully understanding how decisions are made, thereby raising concerns over accountability, fairness, and ethical use. Current explainable AI (XAI) models attempt to address this issue but often fall short of delivering clarity or gaining public confidence. The absence of a unified framework that comprehensively addresses the multifaceted aspects of trust in AI leaves a critical gap in both research and application. Therefore, this research seeks to investigate the root causes of distrust in AI and propose a practical, validated framework that enhances transparency and builds confidence in AI systems across various domains.

1.4 Research Objectives

1. To analyze existing challenges in trust and transparency of AI systems.
2. To assess the effectiveness of current explainable AI frameworks.
3. To develop a comprehensive and practical framework for enhancing trust in AI.
4. To validate the framework across various high-stakes AI application domains.
5. To propose actionable recommendations for developers and policymakers.

1.5 Research Scope

This research focuses on trust challenges in AI within high-stakes fields such as healthcare and legal systems. It evaluates current transparency frameworks, identifies their limitations, and proposes a holistic solution through a newly developed framework tested in multiple real-world contexts.

Chapter 2: Literature Review

The literature review explores previous studies on explainable AI, ethical frameworks, and public trust in intelligent systems. Key studies by Zhang & Lu (2021), Berente et al. (2021), and Reddy (2022) highlight the importance and limitations of current transparency efforts. Further, interdisciplinary research provides insights into how AI can align with societal values. These findings underline the necessity of a practical, integrative framework.

Chapter 3: Research Methodology and Outcomes

3.1 Methodology

This research adopts a mixed-methods approach:

- Phase 1: Literature Review using thematic analysis.
- Phase 2: Qualitative Interviews with AI developers and ethicists.
- Phase 3: Quantitative Surveys targeting AI users and professionals.
- Phase 4: Framework Development and Validation with expert feedback.

3.2 Expected Outcomes

The study expects to:

- Identify current gaps in trust and transparency.
- Provide a validated ethical framework.
- Offer actionable recommendations for enhancing trust in AI systems.

3.3 Gantt Chart

A simplified example of the Gantt chart:

Task	Start Date	End Date	Duration (Days)
Literature Review	01/08/2025	15/08/2025	15
Methodology Design	16/08/2025	25/08/2025	10
Data Collection	26/08/2025	10/09/2025	15
Data Analysis	11/09/2025	20/09/2025	10
Framework Validation	21/09/2025	30/09/2025	10

References

1. Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration*.
2. Reddy, S. (2022). Explainability and artificial intelligence in medicine.
3. Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*.
4. Arrieta, A. B. et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, and challenges. *Information Fusion*.