

Enhancing Trust in Artificial Intelligence: A Comprehensive Analysis and Framework

Course Title: Research Methodology in Computer Science

Course Code: AI800

Professor: Dr. Mohammed Anbar

Student ID: 251001992

Student Name: Meshal Manfi Alkhaldi

Hand-in Date: 10 May 2025

Table of Contents

Abstract	3
Chapter 1: Introduction	4
1.1 Overview	4
1.2 Background	5
1.2.1 The Evolution of Artificial Intelligence	5
1.2.2 Trust and Transparency Issues in AI	6
1.2.3 The significance of Trust and Transparency	7
1.2.4 Challenges to Enhancing Trust in AI	7
1.3 Problem statement.....	8
1.4 Research Objectives.....	9
Research Scope	11
Chapter 2: Literature Review.....	12
2.1 Literature Review.....	12
Chapter 3: Research Methodology.....	16
3.1 Research design	16
Phase 1: Systematic Literature Review.....	16
Phase 2: Qualitative Research through stakeholder interviews	17
Phase 3: Quantitative Research via Surveys.....	18
Phase 4: Framework Development and Validation.....	19
Ethical considerations	20
Limitations	21
Expected outcomes	22
Gantt Chart.....	23
References.....	24

Abstract

The rapid advancement of Artificial Intelligence (AI) has introduced transformative capabilities in vital domains such as healthcare, transport, and the judicial system. But despite all its promise, adoption of AI is hampered by a powerful deterrent—trust deficiency among citizens. This research discusses the intricate challenges to trust in AI, centered around concerns regarding transparency, bias, and ethical disorientation regarding "black box" models. Using a mixed-methods approach—systematic literature review, stakeholder interviews, and quantitative surveys—the study outlines the technical, ethical, and societal challenges obstructing trustworthy AI deployment. The study critically assesses existing explainable AI (XAI) techniques and ethical strategies, indicating deficiencies in scalability, interpretability, and stakeholder alignment. Based on these findings, the study develops a novel interdisciplinary trust framework that integrates technical transparency, moral accountability, and user-centred design. The design is validated through pilot deployments in high-stakes domains like medicine, autonomous vehicles, and justice. The study concludes with practical recommendations for AI developers, policymakers, and institutions to guarantee trust and accountability in AI systems. Ultimately, this research closes the gap between AI innovation and societal embrace by embracing responsible and open AI practice.

Chapter 1: Introduction

1.1 Overview

AI has become prominent in enhancing various sectors, including health, transport, learning institutions, and judicial systems. From effectively diagnosing diseases to driving cars independently, AI can significantly improve efficiency and effectiveness. However, the swift advancement of AI technologies has left society with a new, mostly unforeseen problem. There are a few major concerns, among which the general public's distrust of AI is one of the most significant and potentially devastating factors that may harm AI systems' effectiveness in the future. Trust in AI is a rather complex concept that includes such components as believability, explainability, responsibility, non-biased, and ethically acceptable. In the present age, especially in deep learning models, the reasoning through which an AI concludes cannot be easily understood, hence the name 'black box AI systems' (Adadi & Berrada, 2018). Such opacity benefits no one, and it equally harms end users as much as it creates difficulties for developers, regulators, and policymakers who hope to try implementing proper AI use.

Trust plays an essential role in high-risk areas because the decisions made would result in serious consequences. For example, AI diagnostic tools in healthcare need to garner physician and patient trust to be adopted, yet trust can be eroded due to a lack of explainability (Holzinger et al., 2019). It is the same with Self-Driving cars, where public trust depends on comprehending how the AI processes traffic, whilst with the justice system, data decisions from the AI are questionable and unfair, and they are biased (Pawel et al., 2021). These tend to be accompanied by ethical issues of using AI, including making sure it is moral to make decisions for humans, and legal matters, as more regulatory authorities worldwide call for more openness and

responsibility. Lack of trust affects planning or acceptance by the user and creates a barrier for AI to be implemented in crucial applications.

The current literature shows a need to draw attention to these concerns. Researchers propose the use of explanations for AI (XAI) to fill the gap (Guidotti et al., 2018), as well as discuss trust as a multi-faceted concept that implies technical, ethical, and societal approaches (Pawlowski-Schonberger et al., 2021). However, challenges are looming over current and future solutions, including XAI frameworks and solutions: While incorporated efforts may be helpful to, other methods prove less effective when scaled or fail to align stakeholders' goals (Miller, 2019). This research arises from the pressing need to find ways to ensure the transition between advanced AI technologies and society's need for reliable systems. It provides a holistic approach to developing trust in AI that brings together the technical, ethical and practical solutions according to the development and usage of AI, addressing developers, users, ethicists and policymakers. With this guide, it is possible to promote a trustworthy AI environment, promoting accountability and compliance to get the most out of AI across the vast use cases.

1.2 Background

1.2.1 The Evolution of Artificial Intelligence

The history of AI is briefly as follows: It can be traced back to the mid-twentieth century, famous works of intelligence machines include those of Alan Turing. Artificial Intelligence was at one point in time a strictly theoretical concept which allowed for simple algorithms, however, over the decades, the growth in computational power, the increasing prevalence of big data and the advancement of machine learning and most importantly the deep learning has caused a revolution in the possibilities of application of AI. After the optimistic periods of the 1980s and

the early 1990s, there were what some have termed "AI winters," in which the efforts proved more stagnant than advancements, thus cutting back funding. However, AI rebounded in the twenty-first century due to such discoveries as convolutional neural networks and reinforcement learning. Nowadays, AI outperforms humans in several fields such as image recognition, AI natural language processing and strategic game AI like AlphaGo, as Vinuesa et al. (2020) pointed out. Such advancements have placed the AI at the epicenter of society with uses in health care diagnosis, automobile driving and individualized learning.

1.2.2 Trust and Transparency Issues in AI

Nevertheless, it is essential to mention that the AI's extensive use has shed light on crucial issues concerning trust and transparency. There are several dimensions to the trustworthiness of AI: reliability, safety, fairness of the results, and explainability (Pawel et al., 2021). Unfortunately, building such models, intense neural networks, leads to the problem of interpretability, whereby it is difficult to understand how the decision was arrived at, which is famously known as the black box problem (Guidotti et al., 2018). This lack of interpretability may also confuse even a developer in understanding how AI reaches an inevitable conclusion, especially in sensitive areas such as the diagnosis of diseases, accounting, or the law courts. For example, in healthcare, a clinician cannot trust an AI model if the latter is unable to explain a given recommendation (Holzinger et al., 2019); similarly, in legal systems, low interpretability may reinforce existing bias (Pawel et al., 2021). These factors not only pose a problem in terms of user acceptance but also raise questions of responsibility and consequences to society.

1.2.3 The significance of Trust and Transparency

Trust and transparency should be its core values to ensure that AI is safe and effective in the long term. In essence, for people to embrace AI systems, they must have confidence in the integrity and ethical models embraced by the AI systems by individuals and institutions (Miller, 2019). For high-risk environments like self-driving cars, acceptance is critical due to the likely consequences of capricious actions from such systems (Vinuesa et al., 2020). Likewise, in the medical environment, only fully transparent AI solutions should directly interact with patients or clinicians, essential for building trust and adherence to evidence-based outcomes (Holzinger et al., 2019). Apart from user acceptance, transparency has become a legal and ethical requirement. Some regulatory frameworks, including the GDPR of the European Union, have focused on the concept of 'right to explanation' for automated decision-making, thus calling for interpretable artificial intelligence systems (Pawel et al., 2021). This brings forth the issue that AI will not be able to deliver on its promise of driving innovation and improving societies without trust.

1.2.4 Challenges to Enhancing Trust in AI

While there are mainly technical challenges on the way to becoming trustworthy AI, there is also an array of ethical considerations. On the technical level, there is still a gap in creating interpretable and explainable AI models. Methods like feature importance scores or decision trees focus on explaining the AI decisions and results, but come with a trade-off of reduced accuracy and applicability across multiple problems (Lundberg & Lee, 2017). From an ethical perspective, and to meet human virtues that entail fairness, justice and non-maleficence, Beck et al. (2021) and Pawel et al. (2021) agree that the AI structures must be framed with cultural and contextual considerations. These challenges are further made worse by the fact that the

advancement in the development of AI is speedy, and the regulation of its use is a work in progress. Moreover, the variety in notions of the stakeholders' self-identifying as developers, users, or policymakers requires frameworks that consider technical possibilities and public expectations, which is the purpose of the research (Miller, 2019).

1.3 Problem statement

While AI is present in fields such as care and automobiles, there is a huge issue that no one seems to talk about. Explainability and trust are mainly due to the high opaqueness of today's AI systems, which are 'black box' systems. Professionals argue that the "black box" problem erodes accountability and ethical measures and intensifies the stakeholders' distrust (Adadi & Berrada, 2018). Current XAI has few approaches, some of which have scalability issues and, more importantly, validation. This lack of transparency is not only an issue from the user's point of view, but also in many fields like medical diagnosis, cases in court, etc. These issues are only aggravated by a lack of framework science and its absence of support from all stakeholders. Therefore, there is a great need to enhance trust in the AI system technically, ethically, and socially.

However, increasing the trust in AI goes beyond just addressing the application's vulnerability and poses other questions regarding society. IT expectations can be negative when AI systems lack clarity in their decision making, and this creates scepticism from the end users, considering the technologies as risky or even dangerous (Miller, 2019). This lack of trust may lead to opposition to the deployment of AI, especially in areas that require people's reliance on the solution, such as self-driving cars or AI in education (Vinuesa et al., 2020). The plurality of technical stakeholder interests for developers, usability for users, and ethics for the regulators further compounds this. A lack of a coherent approach towards these diverse narratives

undermines the capacity of AI to produce inclusive value and thereby underlines a need for an elevated and more sensible approach to trust building.

This is especially so given that we are experiencing rising trends of AI uptake and deployment in organizations and the absence of best practices that can enable the development of trust in organizations. While these algorithms are being incorporated into decision-making, the absence of explainability becomes dangerous as it creates the potential of creating a biased system or of compromising the privacy of users (Pawel et al., 2021). Most of the existing solutions, including post-hoc explanation methods such as LIME or SHAP, do not give extensive information on how the model behaves and may not fully meet the explainability requirements, more so for complex models (Ribeiro et al., 2016). This creates a general area for which a new framework can be developed to address the technical issues of recommendation systems. Still, it does not leave the ethical questions aside while aiming to establish trust in multiple artificial Intelligence applications.

1.4 Research Objectives

The objectives of this research include:

1. Investigate barriers to Trust in AI applications:
 - Discuss how AI models suffer from limitations like interpretability and explainability, while ethical concerns of bias, and fairness also come to the foreground when AI is involved in areas like healthcare, automobile, and judicial systems.
 - Evaluate the implications for society and the regulatory system of these barriers on stakeholder confidence.
2. Critically review existing Trust-enhancing approaches:

- Critically discuss the existing approaches in explainable AI (XAI) and the features of the ethical standards to enhance the degree of explainability and trustworthiness.
 - Potential accompanying weaknesses include loopholes which may pertain to scalability, usability, and a lack of coherence with the needs of many stakeholders.
3. Formulate an interdisciplinary Trust framework:
- Using the barrier analysis and solution review, establish a new framework that addresses technical explainability, ethical principles, and stakeholder needs.
 - Likewise, ensure that the framework can be implemented in different subjects related to artificial Intelligence and always consider users' trust.
4. Test and Refine the Framework in Real-World Scenarios:
- It is necessary to carry out pilot studies concerning healthcare, autonomous vehicles, and judicial use of the proposed framework to determine the approach's feasibility in real-life scenarios.
 - Engage stakeholders to collect feedback that will enable the enhancement of the proposed framework for future use.
5. Disseminate Practical guidelines for Trustworthy AI:
- Provide an action plan for implementing the proposed framework to the AI developers, policymakers, and organizations.
 - Engage in research and technological progress of ethical AI to disseminate information and encourage the widespread implementation of trust-building processes among the academic and corporate worlds.

Research Scope

- **Exploration of Trust Barriers:** Investigating technical, ethical, and societal obstacles to trust in AI, focusing on high-stakes domains such as healthcare, autonomous vehicles, and judicial systems, to understand their impact on stakeholder confidence.
- **Assessment of Trust-Enhancing Methods:** Evaluating existing explainable AI (XAI) techniques and ethical guidelines to identify their strengths, limitations, and gaps in addressing scalability and stakeholder needs.
- **Development of a Trust Framework:** Creating an interdisciplinary framework that integrates technical explainability, ethical principles, and stakeholder requirements, ensuring adaptability across diverse AI applications.
- **Practical Validation and Refinement:** Testing the framework through pilot implementations in healthcare, autonomous vehicles, and judicial contexts, refining it based on stakeholder feedback to ensure real-world applicability.
- **Guideline Formulation and Dissemination:** Producing actionable recommendations for stakeholders and contributing to academic and industry discourse on trustworthy AI practices.

To address the aforementioned research objectives, this scope is designed to directly link to identifying trust barriers, assessing the solutions, developing and implementing an appropriate framework, and making available the best guidelines that can be used. This approach is both specific and wide-ranging, as it directly targets all the technical, ethical, and social aspects of trust in artificial Intelligence. Thus, the project's scope ensures that high-stakes domains and multiple stakeholders are engaged in developing a comprehensive, versatile framework that will be helpful in the progress of trustworthy AI.

Chapter 2: Literature Review

2.1 Literature Review

Accelerated advancement of Artificial Intelligence has rendered the topic of trust a critical subject, especially because many artificial Intelligence systems are opaque. This literature review aggregates the most recent research on trust and transparency in AI, including explainable AI (XAI), ethical discussions, and the discussions of their impact on high-stakes applications. This part critically analyses major studies to pick out voids of current approaches and the need for a holistic, stakeholder-aligned framework to support trust in AI.

In their foundational overview of explainable AI (XAI), Adadi and Berrada (2018) focus on the need to reduce the "black box" problem that is inherent to complex AI, including deep neural networks. They introduce a taxonomy of XAI whereby XAI techniques are divided into global and local explanations, post-hoc and intrinsic interpretability, and model-specific and model-agnostic methods (Adadi & Berrada, 2018). Although their work emphasizes the importance of user-centric explanations, there is a shortage of empirical validations in diverse domains, which restricts their practical utility. For example, how these techniques hold up in high-stakes scenarios such as healthcare or judicial systems, where trust is paramount, is not discussed. This gap implies the need for domain-specific evaluation to ensure that different stakeholder needs can be aligned with XAI methods.

Similarly, Guidotti et al. (2018) extensively review techniques for explaining black-box models and propose a unified workflow that combines approaches, such as LIME, SHAP and rule-based explanations (Guidotti et al., 2018). Conceptualized in a very systematic manner, their analysis is helpful because of the contrast between the various explanation options that they discuss. Still, it remains quite theoretical, with insufficient elaboration on situated challenges that would arise in

real-world applications. For instance, the speed that practices such as SHAP take to compute can make scalability difficult in large-scale applications (as also observed by Lundberg & Lee, 2017). Guidotti et al. (2018) note that explanations need to maintain fidelity (accuracy to the model) yet be understandable to an end user (interpretability). Still, they do not discuss the implications for fidelity-interpretability trade-off in areas of high trust, where practical testing, not experimentation and feedback from stakeholders, becomes essential.

Lundberg and Lee (2017) present SHAP (SHapley Additive exPlanations), which combines multiple explanation methods into one approach: consistent and interpretable feature importance scores (Lundberg & Lee, 2017). SHAP's rationale is based on the theoretical understanding of game theory: it allows for the fair attribution of model predictions to input features. The authors show how useful it can be in high-stakes decision-making, such as medical diagnostics, but identify its intensive computation as a barrier to adoption. This limitation points to a more general issue in XAI. The efficient yet accurate methods of developing. Moreover, Lundberg and Lee (2017) do not consider how the SHAP explanations would be compatible with regulatory requirements, for instance, GDPR's "right to explanation"; thus, it is helpful if there are frameworks which incorporate legal and ethical sides.

Ribeiro et al. (2016) introduce LIME (Local Interpretable Model-agnostic Explanations), which attempts to locally interpolate the complex model behavior with less complex and more interpretable models (Ribeiro et al., 2016). The model-agnostic nature of LIME makes it flexible but also interferes with complex models in which the authors admit inconsistencies of local approximations. Though LIME has been deployed in areas like text classification, few studies have been conducted on its performance in high-stakes regions (such as autonomous vehicles).

This constraint highlights the need for resilient XAI approaches in different contexts and has the potential for trust-building with non-technical users, including patients or policymakers.

Ethical considerations are integral to trustworthy AI beyond technical solutions. According to Pawel et al. (2021), ethical challenges of AI include fairness, accountability, and transparency, which are key to trust (Pawel et al., 2021). They argue that technical explainability cannot be done without focusing on biases and any alignment to social values. Their appeal for cooperation with stakeholders is convincing, although the research lacks concrete guidelines on applying the principles of ethics in practice. For example, they do not describe ways to balance transparency with privacy concerns, an essential tension in healthcare applications. This gap calls for frameworks that merge ethical and technical perspectives with the need for AI systems to be interpretable and socially responsible.

Holzinger et al. (2019) concentrate on explainable AI in healthcare, proposing hybrid models that couple deep learning fidelity with simple models' readability. Their work is especially applicable to high-stakes domains where clinician trust is imperative for AI to be adopted. However, they focus on medical applications, which makes generalization of their findings difficult, and they do not delve into scalability pitfalls in resource-constrained settings. This domain-specificity implies the necessity of more general frameworks which could be applied to different contexts, e.g., the judicial system or autonomous vehicles, in which the trust dynamics are different.

Miller (2019) takes an acknowledgement of trust from a cognitive viewpoint; the trustworthiness of AI explanations needs to fit with human ways of reasoning (Miller, 2019). Relieving himself of the formula, he focuses on social science insights when he stresses that explanations should be contrastive – that is, explanatory of why one rather than another result held — and selective in their designation of essential factors. Although fascinating on paper, this

theoretical framework lacks practical techniques for its implementation, in particular, by non-experts, who may have trouble understanding the nitty-gritty of the technical explanations.

Miller's work reinforces the need for a user-centred design approach in XAI. Still, additional work is required to translate these principles into usable tools supporting stakeholder trust-building.

Using the work of Vinuesa et al (2020), this paper analyses AI's contribution to achieving the sustainable development goals and finds that transparency issues limit the potential social implications of AI (Vinuesa et al., 2020). They mention where AI could play a role; for example, in the better use of resources, but also warn that such opaque systems may undermine overall public trust, especially in policy-sensitive ones. Their considerable research presents no deep solutions for trust development, but relatively high-level recommendations. This gap strengthens the argument in favor of specialized frameworks that will address both technical and societal barriers to trust and help integrate it into different sectors responsibly.

These studies demonstrate the intricacy of building trust in AI and call for an integrated technical, ethical, and socio-cultural response. Although XAI methods such as LIME and SHAP have attractive prospects for transparency, there are substantial concerns about their scalability, computer performance, and domain-specific applicability. Ethical considerations of fairness and responsibility are essential, though often vocally absent in concrete implementation strategies. In addition, the literature indicates that relatively few stakeholder-inclusive frameworks consider the needs of developers, users, and regulators. This research seeks to bridge these gaps by providing an integrated, validated framework incorporating technical explainability, ethical principles, and stakeholder input to create faith in AI at high-stakes applications.

Chapter 3: Research Methodology

3.1 Research design

A mixed-method approach is thus used in this study to accomplish the research objectives, as both qualitative and quantitative methods are implemented to have a holistic understanding of trust in AI. Constituting four phases, the methodology is planned in such a way that it has to address particular goals and add on to what was already found in the previous stage. Below, the description of every stage of the research process is presented with objectives, methods, tools, participant details and analytical approaches.

Phase 1: Systematic Literature Review

- **Objective:** To draw a landscape of trust and transparency in AI, marking the essential challenges and gaps in existing solutions and current approaches.
- **Method:** The literature search will be carried out systematically to adhere to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. Peer-reviewed articles published between 2015 and 2025 filed in academic databases such as Google Scholar, PubMed, Scopus, and IEEE Xplore will be sought. Collocations of "trust in AI", "explainable AI", "ethical AI", "transparency" and domain-specific keywords (e.g. "healthcare AI", "autonomous vehicles") will be used as search terms. Inclusion criteria will favor studies that tackle trust, transparency or XAI in high-stakes domains, whereas non-English papers and non-peer-reviewed sources will be eliminated. Tools: Reference management software (Zotero, etc.) will be used to source materials, and NVivo will support the thematic analysis.

- **Analysis:** Thematic analysis of findings shall be used to divide findings into technical, ethical and societal themes and identify designs within and gaps in the literature. A narrative synthesis will condense central learnings that will guide the next steps.
- **Duration:** This stage will last from June 1, 2025, through July 21, 2025 and will provide a duration to collect and analyse the data adequately.

Phase 2: Qualitative Research through stakeholder interviews

- **Objective:** To get closer insights into trust barriers from the views of various players, such as AI developers, ethicists, and end users.
- **Method:** Approximately 20-25 participants will then be interviewed using a semi-structured approach, purposive sampling being used to ensure a representative sample from across the stakeholder groups. Participants will consist of 8-10 AI developers (with areas of expertise in XAI or high-stakes AI applications), 5-7 ethicists (in AI ethics), and 7-8 end users (e.g. clinicians, judicial professionals or consumers of AI systems). Virtual interviews on Zoom, where each interview will take 45-60 minutes and be based on open-ended questions related to perceptions of trust, transparency challenges, and solutions that are wanted. Questions will be individualized to each group (developers will be asked about technical barriers, while end-users will focus on usability).
- **Tools:** Audios will be transcribed with Otter.ai, and NVivo will be used in qualitative coding. Analysis: Qualitative content analysis will be used, based on the approach proposed by Braun and Clarke (2006, thematic analysis type. In the first step, initial coding will identify repeating themes (opacity, bias, regulatory concerns), and then

iterative refinement will be used to create a coding framework. Triangulation with literary review findings will increase validity.

- **Duration:** This phase will last from August 1, 2025, to October 1, 2025, giving plenty of time for us to recruit, interview and analyse the results.

Phase 3: Quantitative Research via Surveys

- **Objective:** To assess the prevalence of trust concerns and perceived usefulness of existing XAI and transparency solutions in a larger population.
- **Method:** We will distribute a cross-sectional survey to approximately 150-200 AI professionals, including developers, researchers, and industry practitioners, who will be invited to respond via professional networks (e.g., LinkedIn, AI conferences) and snowball sampling. The survey will employ a 5-point Likert scale to quantify perceptions of trust barriers (e.g., "How much does model transparency get in the way of trust in AI?") and solution effectiveness (e.g., "How effective is SHAP at building user trust?"). Further questions will gather demographic information (e.g., years of experience, subject matter expertise) and open-ended fields for qualitative feedback. The survey will be administered on Qualtrics and sent electronically.
- **Tools:** Statistical analysis will be conducted using SPSS or R, and survey distribution and data collection will be accomplished using Qualtrics.
- **Analysis:** Response will be summarised quantitatively using descriptive statistics (e.g., frequencies, means), and ANOVA or regression analysis will be used for inferential statistics to examine relations among variables (e.g., experience level and solution perceived impact). Open-ended responses will be coded thematically to augment

quantitative findings. Reliability will be determined by applying Cronbach's alpha to scale items.

- **Duration:** This phase will be conducted from October 1, 2025, through December 12, 2025, at the same time as Phase 2 in order to align data collection schedules.

Phase 4: Framework Development and Validation

- **Objective:** To develop a trust-enhancing framework based on prior findings and validate its applicability in real-world AI applications.
- **Method:** Findings from Phases 1-3 will be synthesised to design a novel framework, integrating technical (e.g., XAI methods), ethical (e.g., fairness principles), and stakeholder (e.g., user-centric design) components. The framework will be iteratively developed through workshops with 10-12 experts (including AI researchers, ethicists, and industry practitioners), conducted virtually or in person, to refine its structure and components. Validation will involve pilot testing in three domains: healthcare (e.g., a diagnostic AI tool), autonomous vehicles (e.g., a navigation system), and judicial systems (e.g., a risk assessment tool). Pilot tests will involve deploying the framework in simulated or controlled environments, with 5-7 stakeholders per domain (e.g., clinicians, engineers, judges) providing feedback via structured questionnaires and focus groups. Feedback will assess the framework's usability, effectiveness, and adaptability.
- **Tools:** Prototyping tools (e.g., Figma) will support framework visualisation, and NVivo will analyse qualitative feedback. Statistical tools (e.g., R) will evaluate questionnaire data.

- **Analysis:** Quantitative feedback will be analysed using descriptive statistics and paired t-tests to assess perceived trust pre- and post-framework implementation improvements. Qualitative feedback will be thematically analysed to identify refinement areas. Iterative revisions will ensure the framework's robustness.
- **Duration:** This phase will span from December 16, 2025, to January 16, 2026, allowing time for framework development, pilot testing, and refinement.

Ethical considerations

The study upholds ethical standards to obtain participant privacy, safety, and trust at all phases. Informed consent will be administered to all participants, such as survey respondents, interviewees, and stakeholders of the pilot test, through clear, simple consent forms explaining the purpose of the study, procedures, risks, and benefits. Participants will be notified of their right to withdraw at any time without penalty, and contact information will be provided for research team and ethics board questions or concerns. For privacy assurance, all information (e.g., interview audio, survey data, pilot test results) will be de-identified using unique identifiers, and personally identifying information (e.g., names, contact information) will be kept separately in encrypted databases that can be accessed by the research team only. Data will be kept on university-approved servers with multi-factor security, and access will be limited to authorised personnel only.

Possible ethical risks are psychological harm from exploring sensitive issues (e.g., AI bias in criminal justice systems) or inadvertent revelation of trade-secret data by industry respondents. To balance these, interview questions will be framed to avoid traumatic issues, and respondents will be told to share only non-confidential data. Pilot testing will utilise controlled or simulated settings to prevent harm to the actual world, specifically in high-stakes areas such as judicial

systems or healthcare. Approval from the university's Institutional Review Board (IRB) will be obtained before collecting data, and comprehensive protocols will be submitted for each step. There will be regular ethics audits for adherence to GDPR and other applicable laws to the extent of data protection and the "right to explanation." To mitigate power imbalance, particularly with end-users who might consider themselves less experienced in AI, simple explanations in plain language and voluntary debriefing sessions will be provided to promote understanding and comfort. These actions seek to promote beneficence, non-maleficence, autonomy, and justice in research practice.

Limitations

The study has several potential limiting factors that might affect its findings and relevance. First, the sample size for qualitative interviews at 20 to 25 respondents and surveys on 150 to 200 subjects could limit the diversity of opinions, particularly if specific stakeholder groups like policymakers are inadequately covered. This might result in outcomes that fail to capture the entire range of barriers to trust among different AI fields. To address this issue, purposive and snowball sampling techniques will be employed to ensure a diverse representation, while triangulation with findings from the literature review will bolster the study's robustness.

Furthermore, concentrating on three high-stakes areas—healthcare, autonomous vehicles, and judicial systems—might limit the framework's relevance to alternative contexts like education or finance. Pilot testing will encompass a variety of use cases within these domains to optimise generalizability, and recommendations for future research will suggest the examination of further sectors.

Third, survey response bias, for instance, social desirability or acquiescence bias, will tend to skew results if respondents sense pressure to report positive attitudes towards AI solutions.

Tested survey instruments and anonymous answers will be used to mitigate the risk, and open-ended questions will allow subtle observation. Fourth, the pilot test period may face logistical challenges, such as limited access to real AI systems or stakeholder availability, which may undermine the level of validation. Controlled environments and flexible scheduling will reduce these constraints, but findings will still be based on artificial conditions rather than real-world complexity. Finally, the rapid pace of AI technology evolution and regulation over the study period (2024-2025) may render some findings obsolete. Ongoing update of literature and specialist input during framework development will ensure that it remains relevant. Such restrictions, though vital, are known to guide the interpretation of results and inform future research directions.

Expected outcomes

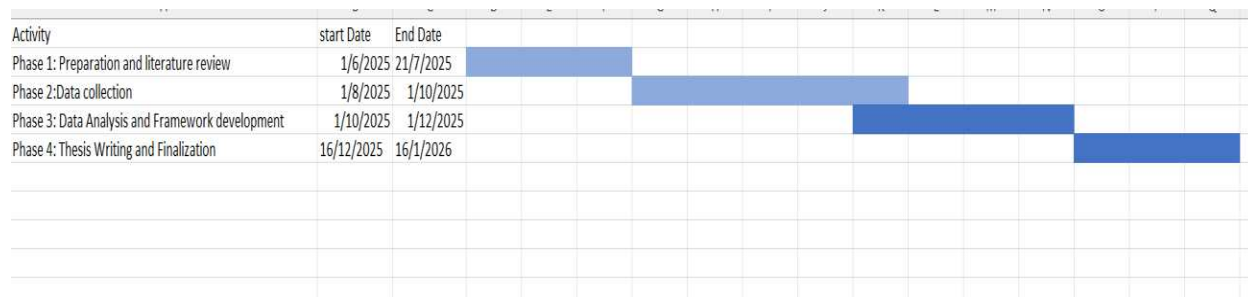
The study is designed to yield several impactful outcomes that advance the understanding and practice of trustworthy AI:

- **In-Depth Analysis of Trust Barriers:** The research will provide a report of thorough literature- and stakeholder-informed technical (e.g., model opaqueness), ethical (e.g., bias, fairness), and societal (e.g., public distrust) trust obstacles in AI. The analysis will provide a qualitative understanding of trust mechanisms in high-stakes environments, which can guide researchers and practitioners who seek to address similar problems.
- **Analysis of Current Solutions:** A critical analysis of existing XAI methods (e.g., LIME, SHAP) and ethical guidelines will establish their strengths, weaknesses, and limitations regarding scalability, usability, and adherence. This analysis will enrich the XAI knowledge with empirical knowledge about solution performance, guiding developers on which tools to use for a specific application.

- **Novel Trust-Enhancing Framework:** The study will deliver a validated, stakeholder-matched framework uniting technical explainability, ethical values, and human-centred design. The framework will be versatile in multiple AI contexts, offering developers, policymakers, and organization's a real-world tool to enhance trust.
- **Actionable Recommendations:** Practical recommendations will be made available to stakeholders, like a step-by-step deployment strategy for the framework, explainability best practices, and compliance ethics checklists. These recommendations will facilitate industry professionals embracing trusted AI practices and assist policymakers in writing informed legislation by bridging the research-to-application gap.
- **Academic and Industry Impacts:** The research will enhance the discourse on ethical AI in peer-reviewed journals, conference papers, and industry seminars, building broader awareness and adoption of trust-enhancing practices. Engaging diverse stakeholders will spur cross-disciplinary dialogue, inspiring future research into large-scale XAI methodologies and ethics. The findings will also inform public policy debates on AI regulation, resulting in societal trust in AI systems by 2030.

Gantt Chart

Below is the Gantt Chart for the research.



References

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial Intelligence (XAI). *IEEE Access*, 6, 52138-52160.
<https://doi.org/10.1109/ACCESS.2018.2870052>
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 1-42. <https://doi.org/10.1145/3236009>
- Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2019). What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923*.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774.
- Miller, T. (2019). Explanation in artificial Intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1-38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Pawel, M., Brocki, L., & Chung, N. C. (2021). The ethics of trustworthy AI: Challenges and opportunities. *Frontiers in Artificial Intelligence*, 4, 672387.
<https://doi.org/10.3389/frai.2021.672387>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.
<https://doi.org/10.1145/2939672.2939778>
- Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., ... & Nerini, F. F. (2020). The role of artificial Intelligence in achieving the Sustainable Development

Goals. *Nature Communications*, 11(1), 233. <https://doi.org/10.1038/s41467-019-14108-y>