

Enhancing Trust in Artificial Intelligence: A Comprehensive Analysis and Framework

Table of Contents

Abstract	1
Introduction.....	2
Background	3
The Evolution of Artificial Intelligence	3
Trust and Transparency Issues in AI	3
The Significance of Trust and Transparency	3
Challenges to Enhancing Trust in AI	4
Problem Statement	4
Research Objectives	5
Research scope	6
Literature review	8
Research Methodology	15
Phase 1: Literature Review	15
Phase 2: Qualitative Research	15
Phase 3: Quantitative Research	16
Phase 4: Framework Development and Validation.....	16
Ethical Considerations.....	17
Limitations	17
Expected Outcomes.....	17
Gantt chart	19
References	20

Abstract

The rapid integration of Artificial Intelligence (AI) into critical sectors of society underscores an emergent need: the establishment of trust in AI systems. This thesis examines the multifaceted nature of trust in AI, addressing the opacity of advanced AI models and the consequent ethical, legal, and societal concerns. Drawing on a mixed-methods approach that combines a systematic literature review, qualitative interviews, and quantitative analysis, this research aims to dissect the complex relationship between AI and its stakeholders. This study's foundation is built upon an in-depth review of current challenges and proposed solutions within the realm of explainable AI (XAI), focusing particularly on high-stakes domains such as healthcare, autonomous vehicles, and judicial decision-making.

The principal contribution of this work is the development of a novel framework designed to enhance trust in AI. This framework, informed by interdisciplinary insights and validated by both AI experts and laypersons, introduces pragmatic guidelines that address the nuances of ethical AI deployment. Additionally, the research provides a comprehensive assessment of AI's potential societal impact by 2030, highlighting the importance of scientifically grounded portrayals of AI's capabilities and limitations.

Keywords: Artificial Intelligence, Trust, Explainable AI, Ethical AI Framework, Stakeholder Analysis

Introduction

In an era where artificial intelligence (AI) seamlessly integrates into every facet of human life, from healthcare diagnostics to automated driving and personalized education, the promise of AI to enhance efficiency, decision-making, and innovation is undeniable. However, the accelerated adoption of AI systems brings forth a critical challenge that threatens to undermine these advancements: the lack of trust in AI technologies. Trust, a multifaceted and indispensable component in human-AI interaction, is compromised by issues of transparency, accountability, and ethical considerations. As AI systems become more complex, their "black box" nature exacerbates these issues, rendering their decision-making processes opaque and difficult to understand even for their creators.

The significance of trust in AI is underscored in various domains, particularly in fields where the stakes are high, such as medicine, autonomous vehicles, and legal adjudication. In these areas, the inability to comprehend and predict AI behaviors not only hampers user acceptance but also raises ethical and legal concerns. This problem is highlighted in the existing literature, with scholars like Reddy (2022) and Zhang and Lu (2021) articulating the challenges posed by the opaque nature of advanced AI models. These challenges underscore the urgent need for frameworks that enhance the transparency, explainability, and ultimately, the trustworthiness of AI systems.

This research is motivated by the critical gap between the current capabilities of AI technologies and the societal need for transparent, ethical, and trustworthy AI systems. It seeks to address this gap by proposing a comprehensive framework that enhances trust in AI across various dimensions: technological, ethical, and practical. This framework aims to align with the needs of stakeholders, including developers, users, and policymakers, fostering a more ethical and transparent AI landscape.

Background

The Evolution of Artificial Intelligence

Artificial Intelligence (AI) has transitioned from the realms of science fiction to an integral component of modern society. Its roots can be traced back to the mid-20th century when the concept of "machine intelligence" was first theorized. Since then, AI has undergone several waves of optimism and skepticism, leading to periods of rapid development and so-called "AI winters." Today, we find ourselves amidst an AI renaissance, fueled by advancements in computational power, the availability of big data, and breakthroughs in machine learning algorithms, particularly deep learning. These developments have enabled AI to exceed human performance in specific tasks, ranging from image recognition to strategic game playing.

Trust and Transparency Issues in AI

As AI applications proliferate, they bring to light critical issues related to trust and transparency. Trust in AI is multifaceted, encompassing reliability, safety, fairness, and the ability to understand and predict AI behavior. However, the opacity of complex AI models, especially those based on deep learning, poses significant challenges to achieving transparency. This opacity, often referred to as the "black box" problem, means that even the engineers who design AI systems may not fully understand how decisions are made. This lack of transparency is problematic in any context but is particularly critical in high-stakes areas such as healthcare, finance, and legal adjudication, where decisions can have profound impacts on human lives.

The Significance of Trust and Transparency

The significance of trust and transparency in AI cannot be overstated. For AI to reach its full potential, it must be adopted and accepted by end-users, and this acceptance is heavily contingent upon trust. Trustworthy AI is essential not only for individual

user comfort and safety but also for the societal adoption of AI technologies. Furthermore, as regulatory bodies around the world begin to scrutinize AI more closely, the ability to demonstrate the transparency and ethical grounding of AI systems becomes a legal and ethical imperative.

Challenges to Enhancing Trust in AI

Enhancing trust in AI involves addressing both technical and ethical challenges. Technically, developing methods to make AI systems more interpretable and explainable remains a significant research focus. Ethically, ensuring that AI systems make decisions in a manner that aligns with human values and ethics is paramount. These challenges are compounded by the rapid pace of AI development, which often outstrips the slower processes of ethical consideration and regulation.

Problem Statement

In the rapidly evolving landscape of artificial intelligence (AI), the lack of transparency and trust in AI systems, particularly in domains like medicine, hinders their widespread adoption. The black-box nature of many advanced AI models, as highlighted by Reddy (2022) and Zhang and Lu (2021), poses significant challenges in ensuring accountability, ethical usage, and understanding of decision-making processes. Existing solutions, such as explainable AI (XAI) frameworks, have faced criticism for their imperfections (Reddy, 2022) and the need for alternative validation approaches (Ghassemi et al., 2022).

The significance of this problem is underscored by its impact on critical domains like healthcare, where trust and transparency are paramount. As AI increasingly augments or replaces functions of physicians, the inability to explain decisions erodes patient trust (Reddy, 2022). The problem domain extends beyond medicine, encompassing AI management complexities (Berente et al., 2021), the potential of

AI to contribute to sustainable development (Goralski & Tan, 2020), and the broader societal implications outlined in the One Hundred Year Study on Artificial Intelligence (Stone et al., 2022).

This research aims to address the pressing need for enhancing trust in AI by comprehensively analyzing existing challenges, evaluating solutions, and proposing a robust framework aligned with the identified problem challenges.

Research Objectives

1. To comprehensively analyze the existing challenges in trust and transparency of AI:
 - To identify the key challenges and limitations of current AI systems, as outlined in the literature.
 - To understand the ethical, legal, and social implications of opaque AI decision-making processes.
2. To evaluate existing solutions, including explainable AI frameworks:
 - To critically assess the strengths and weaknesses of explainable AI methods in enhancing transparency.
 - To explore alternative validation approaches proposed by scholars such as Ghassemi et al. (2022).
3. To propose a comprehensive framework for enhancing trust in artificial intelligence:
 - To integrate insights from the analysis of existing challenges and evaluations of solutions.

- To develop a framework that addresses the identified problem challenges and aligns with ethical considerations.
4. To validate the proposed framework in diverse AI application domains:
- To assess the generalizability and effectiveness of the proposed framework in various AI contexts.
 - To gather feedback from practitioners, stakeholders, and experts in the field to refine the framework.
5. To provide practical recommendations for the implementation of the proposed framework:
- To offer actionable guidance for organizations, policymakers, and developers to integrate the framework into AI systems.
 - To contribute to the ongoing discourse on responsible and trustworthy AI development and deployment.

Research scope

Analysis of Existing Challenges: Concentrating on identifying and understanding the key issues that currently hinder trust and transparency in AI, with a particular emphasis on high-stakes domains such as healthcare, autonomous vehicles, and judicial decision-making systems.

Evaluation of Current Solutions: Exploring existing solutions, including but not limited to, explainable AI (XAI) frameworks, transparency guidelines, and ethical AI principles. This evaluation seeks to assess the efficacy and limitations of these solutions in addressing the identified challenges.

Development of a New Framework: Proposing a novel framework that integrates insights from the analysis and evaluations. This framework aims to offer practical and ethical guidelines to enhance trustworthiness in AI systems.

Validation in Diverse Contexts: Applying the proposed framework to varied AI application domains to assess its generalizability and effectiveness. This step includes seeking feedback from a range of stakeholders, including AI developers, users, and policymakers.

Literature review

Zhang and Lu's (2021) research, published in the Journal of Industrial Information Integration, provides a pivotal exploration of artificial intelligence (AI), delineating its current state and forecasting future advancements. By methodically reviewing significant AI-related publications, their work highlights AI's comprehensive scope, including its technological progress, applications, and interdisciplinary nature spanning information science, logic, cognitive studies, and more. Their findings underscore AI's potential in enhancing logical reasoning and emotional capacities, possibly surpassing human intelligence. However, the study's limitations include a lack of transparency in article selection criteria and a need for deeper integration of AI's multidisciplinary aspects, alongside a somewhat speculative forecast regarding AI's advancement without solid empirical backing. Despite these critiques, Zhang and Lu's contribution remains notable for its broad overview and forward-looking perspective on AI, offering a valuable point of reference for understanding AI's evolving landscape and its implications for future research.

In their 2021 article in MIS Quarterly, Berente, Gu, Recker, and Santhanam address the nuanced complexities involved in managing artificial intelligence (AI), advocating for a socio-technical framework to navigate AI's autonomy, learning capabilities, and opacity. Their research underscores the importance of interdisciplinary approaches in AI management, proposing the information systems discipline as a cornerstone for guiding the field. Through a diverse methodological approach that includes observational studies, experimental designs, and econometric analysis, the authors illuminate the evolving landscape of AI management. They challenge existing disciplinary norms and highlight the impact of AI on academic review processes, contributing a fresh perspective to the scholarly debate on AI's role within and beyond information systems. However, the article's impact is

somewhat tempered by its limited discussion on specific case studies, calling for a deeper integration of tangible examples to fully articulate the dynamics of effective AI management across disciplines. This study enriches the literature on AI by framing management challenges within the context of AI's unique characteristics and the interdisciplinary effort required to address them.

In their 2020 study published in *The International Journal of Management Education*, Goralski and Tan explore the intersection of artificial intelligence (AI) and sustainable development, particularly its alignment with the United Nations Sustainable Development Goals (SDGs). Through a blend of business strategy and public policy lenses, they analyze three case studies to underscore AI's potential in promoting sustainable economic growth, environmental stewardship, and social inclusiveness. Their findings suggest that AI can significantly contribute to resource management efficiency, innovation, and the facilitation of cross-sector collaborations for sustainability. Moreover, the study delves into the transformative effect of AI on education, indicating shifts in teaching methodologies, learning experiences, and the educator's role. While the paper champions the integration of AI in advancing SDGs, highlighting the need for incentives and partnerships, it also acknowledges its reliance on qualitative assessments and the absence of a detailed exploration of AI's potential adverse effects. Goralski and Tan's work thus presents a nuanced view of AI's contributions to sustainability, advocating for a balanced approach to leveraging technology in achieving global sustainability goals, albeit with a noted gap in empirical validation and a comprehensive discussion on associated challenges.

In their pioneering report for Stanford's "One Hundred Year Study on Artificial Intelligence," Peter Stone et al. (2022) provide a forward-looking analysis on how artificial intelligence (AI) might shape various societal sectors by 2030. This

comprehensive study, drawing on the expertise of a panel of 17 AI specialists, delves into eight critical areas: transportation, healthcare, education, public safety and security, low-resource communities, employment and workplace, and entertainment, aiming to paint a scientifically and technologically realistic picture of AI's current capabilities and its potential impacts. The research's multidisciplinary approach, incorporating insights from AI research, business strategy, and public policy, aims to inform and guide decision-making across industry and government sectors. Key findings suggest a significant influence of AI across the identified domains, highlighting both opportunities and challenges. While the report is celebrated for its broad assessment and the integration of diverse expert perspectives, it faces critique for not detailing its methodological framework, which could affect reproducibility and objectivity due to potential expert biases. Furthermore, the speculative nature of forecasting AI's societal impact by 2030 introduces uncertainties that may compromise the precision of its predictions. This landmark study contributes to ongoing dialogues about AI's societal roles, even as it navigates the inherent difficulties of long-term forecasting in a rapidly evolving technological landscape.

Sandeep Reddy's 2022 study, published in April, scrutinizes the imperative role of explainable artificial intelligence (AI) within the medical sector, focusing on addressing the opacity often associated with AI algorithms. Reddy emphasizes the critical need for transparency in medical decision-making, particularly for ensuring adherence to evidence-based practices, and counters the skepticism towards explainable AI by underlining its significance in fostering trust and reliability in AI-driven medical applications. By engaging in a critical analysis that includes addressing objections raised by Ghassemi and colleagues, Reddy not only champions the necessity of explainable AI from an ethical and patient autonomy perspective but also critiques the reliance on black-box models that may leverage

incorrect or misleading variables, thus potentially jeopardizing patient care. Despite the compelling arguments for explainability, the paper's limitations include a lack of engagement with the technical nuances of explainable AI methods and a somewhat unilateral focus that may overlook the benefits of non-explainable approaches. Nonetheless, Reddy's advocacy for explainable AI in medicine contributes to the broader discourse on the ethical, transparency, and trust imperatives in AI application in healthcare, highlighting the nuanced balance between technological advancement and ethical responsibility.

Thiebes, Lins, and Sunyaev's 2021 article in *Electron Markets* introduces and delineates the concept of Trustworthy Artificial Intelligence (TAI), proposing a framework based on five foundational principles: beneficence, non-maleficence, autonomy, fairness, and explicability. This study underscores the critical importance of trust in the development, implementation, and application of AI, presenting a data-driven research methodology named DaRe4TAI to explore technical and non-technical means for supporting TAI. Highlighting the framework's utility, the authors explore future research trajectories, notably focusing on distributed ledger technology (DLT) as a novel avenue for realizing TAI. Despite its significant contributions to conceptualizing TAI and outlining a structured approach to its study, the paper's limitations include a scarcity of concrete examples or case studies to illustrate the application of TAI principles and a need for a more detailed examination of potential tensions between these principles and various data stages. Nonetheless, Thiebes, Lins, and Sunyaev's work aligns with broader discussions on the ethical, legal, and social implications of AI, contributing a crucial perspective on the necessity of trust and ethical considerations in AI's evolution and integration into society.

Arrieta et al.'s (2020) extensive review on Explainable Artificial Intelligence (XAI) in "Information Fusion" outlines the evolution and critical aspects of making AI systems more interpretable and understandable to humans. They emphasize the significance of XAI in facilitating responsible AI development, addressing the inherent opacity of advanced AI models like deep neural networks. The study proposes a comprehensive framework for XAI, including concepts, taxonomies of explainability approaches, and identifies key opportunities and challenges. Through a detailed analysis, they advocate for XAI as essential in achieving transparency, fairness, and accountability in AI applications across various sectors. Their work critically evaluates current methodologies and suggests directions for future research, highlighting the crucial balance between model complexity and explainability. The article stands as a seminal contribution, aiming to bridge the gap between AI's technical capabilities and the societal imperative for trustworthy, ethical AI systems. This work is instrumental for researchers, providing a foundational understanding of XAI, its importance for responsible AI deployment, and guiding principles for developing understandable AI solutions.

Binns (2018) in his seminal work presented at the Conference on Fairness, Accountability, and Transparency, explores the conceptual overlap between fairness in machine learning (ML) and the principles of political philosophy, offering a nuanced examination of what constitutes 'fairness' in ML algorithms. By dissecting various political and moral philosophies, Binns critically evaluates the existing ML fairness metrics, revealing inherent conflicts and the impossibility of satisfying them simultaneously. His analysis underscores the complexity of operationalizing fairness, challenging the ML community to consider deeper ethical and philosophical foundations in the development and implementation of algorithms. Through a philosophical lens, Binns questions the adequacy of current fairness

approaches in ML, suggesting that a more profound engagement with political philosophy can illuminate the paths toward genuinely fair and responsible AI systems. This study not only broadens the discussion on algorithmic fairness but also highlights the intricate relationship between technology and societal values, urging for a multidisciplinary approach in addressing fairness and discrimination within the rapidly evolving domain of artificial intelligence.

Floridi et al. (2018) in "AI4People—an ethical framework for a good AI society" present a foundational framework aimed at guiding the ethical development and application of Artificial Intelligence (AI) towards fostering a "Good AI Society". Their work emphasizes the dual nature of AI as a potent tool for societal benefit while also acknowledging its potential risks. They articulate this through a synthesis of five ethical principles: beneficence, non-maleficence, autonomy, justice, and explicability, offering a novel approach to ethical AI by integrating the latter principle to ensure intelligibility and accountability. Additionally, the study proposes 20 concrete recommendations aimed at policymakers, researchers, and practitioners to ensure that AI's development aligns with ethical standards, promoting inclusivity, fairness, and transparency. Floridi et al.'s contribution is significant in its comprehensive approach to ethical considerations in AI, aiming to catalyze a multi-stakeholder dialogue to harness AI's potential responsibly while mitigating its risks, thereby laying a robust groundwork for ethical guidelines in the burgeoning field of AI.

Jobin, Ienca, and Vayena (2019) in their study published in *Nature Machine Intelligence* embark on a comprehensive examination of the burgeoning landscape of AI ethics guidelines globally, uncovering both convergence and divergence in the principles underpinning ethical AI development. Their analysis, based on a scoping review of 84 documents issued by a variety of stakeholders, identifies an emerging

consensus around five core ethical principles: transparency, justice and fairness, non-maleficence, responsibility, and privacy. However, their deep dive reveals substantial variation in the interpretation, prioritization, and application of these principles, reflecting the complexity of operationalizing ethics in AI across different contexts. This study critically reflects on the alignment and conflicts among global stakeholders' ethical frameworks for AI, emphasizing the necessity of bridging conceptual gaps and harmonizing implementation strategies to foster truly responsible AI development. The work of Jobin et al. stands as a pivotal reference for understanding the current ethical landscape in AI, highlighting the importance of interdisciplinary dialogue and concerted action to navigate the ethical challenges posed by AI technologies.

Research Methodology

This research adopts a mixed-methods approach to comprehensively analyze the existing challenges and solutions related to trust and transparency in AI, and to develop and validate a comprehensive framework for enhancing trust in AI. This methodology is chosen for its strength in providing a holistic understanding of the research problem by combining the depth of qualitative insights with the breadth of quantitative analysis.

Phase 1: Literature Review

The first phase involves an extensive literature review to identify and synthesize existing research on trust and transparency in AI. This review covers scholarly articles, conference papers, industry reports, and relevant legal and ethical guidelines. The purpose is to map out the current landscape of trust in AI, pinpointing gaps in knowledge and practice that the proposed framework aims to address.

Data Sources: Academic databases (e.g., IEEE Xplore, PubMed, ACM Digital Library) and professional publications.

Selection Criteria: Publications from the last ten years, focusing on trust, transparency, ethical implications, and explainable AI.

Analysis Technique: Thematic analysis to identify recurring themes, challenges, and solutions.

Phase 2: Qualitative Research

The second phase involves qualitative research through semi-structured interviews and case studies. The goal is to gain deeper insights into the experiences and perceptions of stakeholders regarding trust and transparency in AI systems.

Participants: A purposive sample of AI developers, users, ethicists, and policymakers.

Data Collection: Semi-structured interviews conducted remotely or in person, adhering to ethical research standards including informed consent.

Analysis: Qualitative content analysis to extract themes related to trust challenges and effective strategies for enhancing transparency and ethical compliance in AI.

Phase 3: Quantitative Research

This phase complements the qualitative insights with quantitative data, assessing the prevalence of identified challenges and evaluating the effectiveness of existing solutions.

Survey Design: A survey based on the qualitative phase findings, targeting a broader audience of AI stakeholders.

Sample: Random sampling of professionals across key sectors utilizing AI, aiming for statistical significance in the responses.

Data Analysis: Statistical analysis, including descriptive statistics and inferential statistics, to understand the distribution of perceptions and the relationship between variables related to trust in AI.

Phase 4: Framework Development and Validation

Based on insights gathered from the previous phases, a comprehensive framework for enhancing trust in AI will be developed. This framework will be validated through expert reviews and pilot testing in real-world AI applications.

Development: Integration of identified challenges, ethical considerations, and effective transparency strategies into a coherent framework.

Validation: Expert feedback from academics and practitioners in the field of AI and ethics, followed by pilot testing in selected AI projects.

Evaluation Criteria: Clarity, applicability, comprehensiveness, and effectiveness in enhancing trust.

Ethical Considerations

Throughout the research, ethical considerations are paramount. This includes obtaining informed consent from participants, ensuring confidentiality and anonymity, and minimizing any potential harm.

Limitations

This research methodology acknowledges potential limitations, including the representativeness of the sample in qualitative research and the response bias in survey data. Efforts will be made to mitigate these limitations through careful sampling and data triangulation.

Expected Outcomes

Comprehensive Analysis of Trust Challenges in AI: A detailed examination of the current challenges and barriers to trust in AI systems across various domains. This analysis is expected to highlight the multifaceted nature of trust in AI, including issues related to transparency, accountability, ethics, and societal impact.

Evaluation of Existing Solutions: An assessment of current methodologies and frameworks aimed at enhancing AI trustworthiness. This evaluation is anticipated to uncover gaps in existing approaches, particularly in terms of their scalability, applicability across different AI domains, and effectiveness in addressing users' and stakeholders' concerns.

Development of a Novel Framework: The formulation of an innovative, comprehensive framework for enhancing trust in AI. This framework is expected to integrate insights from the literature review, qualitative, and quantitative research phases, providing actionable guidelines for AI developers, policymakers, and users to implement trustworthy AI practices. It aims to address identified gaps and offer a balanced approach to technical and ethical considerations.

Practical Recommendations for Stakeholders: The derivation of practical, actionable recommendations for various stakeholders involved in AI development and deployment. These recommendations will be tailored to address the specific needs and concerns of developers, users, policymakers, and ethicists, promoting a collaborative effort towards the development of more transparent and ethical AI systems.

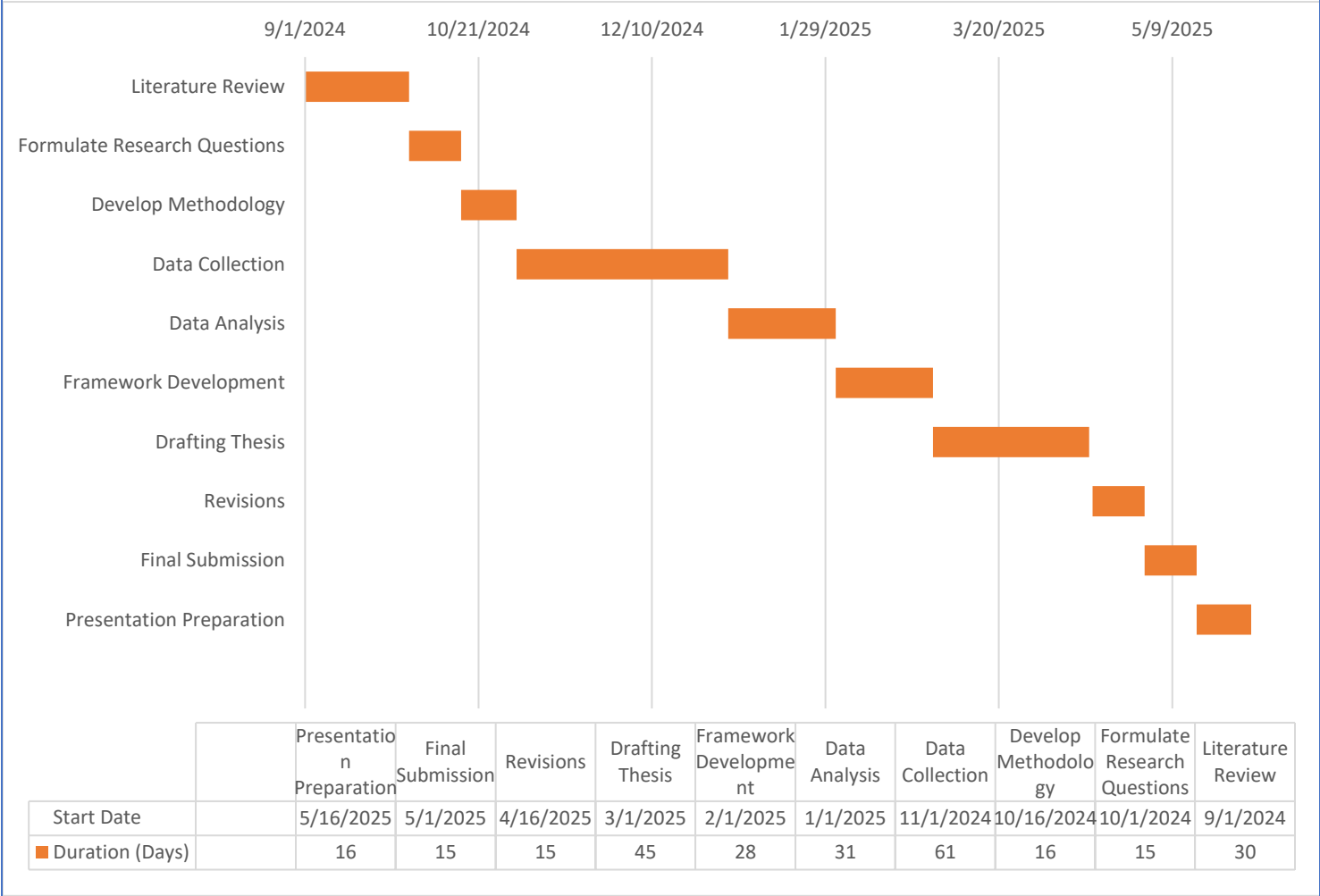
Contribution to Academic and Industry Discourse: The enrichment of academic and industry discourse on trust and ethics in AI. By providing a well-researched and critically analyzed body of work, this research aims to stimulate further discussion, research, and innovation in the field, bridging gaps between theoretical ethics and practical implementation in AI technologies.

Foundation for Future Research: Establishment of a solid foundation for future research on trust in AI. This study is expected to identify new research questions and areas for further investigation, encouraging ongoing exploration of how trust in AI can be enhanced and sustained as technologies evolve.

Enhanced Public Awareness and Engagement: By addressing the complex issue of trust in AI, this research also aims to enhance public awareness and engagement on the topic. Increased understanding among the public regarding the importance of

trust and transparency in AI can foster greater demand for ethical AI practices, influencing industry standards and regulatory policies.

Gantt chart



References

1. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Benetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115. <https://doi.org/10.1016/j.inffus.2019.12.012>
2. Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, 45(3), 1433-1450. <https://doi.org/10.25300/MISQ/2021/16274>
3. Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 149-159. <https://doi.org/10.1145/3278721.3278779>
4. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B. (2018). AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689-707. <https://doi.org/10.1007/s11023-018-9482-5>
5. Goralski, M. A., & Tan, T. K. (2020). Artificial intelligence and sustainable development. *The International Journal of Management Education*, 18(1), 100330. <https://doi.org/10.1016/j.ijme.2019.100330>
6. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389-399. <https://doi.org/10.1038/s42256-019-0088-2>
7. Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania*

Law Review, 165, 633. Retrieved from
<https://www.pennlawreview.com/print/?id=587>

8. Reddy, S. (2022). Explainability and artificial intelligence in medicine. *Published in April, 2022.* [https://doi.org/10.1016/S2589-7500\(22\)00029-2](https://doi.org/10.1016/S2589-7500(22)00029-2)
9. Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M., & Teller, A. (2022). Artificial Intelligence and Life in 2030: The One Hundred Year Study on Artificial Intelligence. *arXiv:2211.06318.*
<https://doi.org/10.48550/arXiv.2211.06318>
10. Thiebes, S., Lins, S., & Sunyaev, A. (2021). Developing a data-driven research framework for trustworthy artificial intelligence. *Electronic Markets*, 31(3), 447-464.
11. Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76-99.
<https://doi.org/10.1093/idpl/ix005>
12. Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration*, 23, 100224. <https://doi.org/10.1016/j.jii.2021.100224>