



Vaasan yliopisto
UNIVERSITY OF VAASA

Antti P. Happonen

**Trustworthy AI in Healthcare Innovation: Ethical Assessment and Project
Management under the EU AI Act**

School of Technology and Innovations
Strategic Project Management Programme

Vaasa 2026

UNIVERSITY OF VAASA**School of Technology and Innovations**

Author:	Antti P. Happonen		
Title of the thesis:	Trustworthy AI in Healthcare Innovation: Ethical Assessment and Project Management under the EU AI Act		
Degree:	Master's Programme in Industrial Engineering and Management		
Degree Programme:	Strategic Project Management Programme		
Supervisor:	Marko Mäkilouko		
Year:	2026	Pages:	132

ABSTRACT: The High-Level Expert Group (HLEG) on Artificial Intelligence (AI), established by the European Commission, defines ethical guidelines for trustworthy AI. These ethical guidelines describe seven key principles that AI developers must follow to ensure the ethical and responsible development of AI. The HLEG guidance has been directly transferred to the EU Artificial Intelligence Act adopted in 2024. Trustworthy AI refers to AI systems that are explainable, fair, interpretable, robust, transparent, safe and secure. The central question in this thesis is how these seven AI HLEG key ethical principles can be integrated into healthcare product innovation and development work and further into project management practices. As AI becomes increasingly integral to both development processes and outcomes, the demand for transparent, structured, and ethically grounded governance has intensified. This thesis addresses this need by examining the suitability of the Z-Inspection® method for assessing AI trustworthiness and by exploring how such evaluations can be integrated into the project management lifecycle. The implications of this thesis extend beyond individual projects. For practitioners, our findings emphasize the need to build AI governance competencies, cultivate responsible AI cultures, and establish cross-functional structures for socio-technical assessment. For policymakers, our research highlights the importance of clearer regulatory guidance, alignment between AI-specific and healthcare-specific requirements, and broader capacity-building efforts. For researchers, this thesis identifies opportunities to further operationalize trustworthy AI assessment, evaluate frameworks such as Z-Inspection®, and study organizational maturity in AI governance. Overall, this thesis demonstrates that integrating trustworthy AI assessment into project management is both feasible and necessary. As AI continues to shape healthcare innovation, organizations must adopt context-sensitive management approaches to ensure safety, transparency, and alignment with societal values. By connecting ethical principles with practical implementation, this work contributes to strategic project management practices that support trustworthy AI innovation. The proposed integration of Z-Inspection® into the project management lifecycle offers a structured pathway for navigating AI adoption while upholding innovation, trustworthiness, and regulatory compliance under the EU legislation.

KEYWORDS: Artificial Intelligence, Ethics, Healthcare, Innovation, Project Management, Regulation (Control), and Trust.

VAASAN YLIOPISTO**Tekniikan ja innovaatiojohtamisen akateeminen yksikkö**

Tekijä:	Antti P. Happonen		
Opinnäytetyön otsikko:	Luotettava tekoäly terveydenhuollon innovaatioissa: eettinen arviointi ja projektinhallinta EU:n tekoälylain nojalla		
Tutkinto:	Tuotantotalouden diplomi-insinööri		
Tutkinto-ohjelma:	Strategisen projektijohtamisen opintosuunta		
Valvoja:	Marko Mäkilouko		
Vuosi:	2026	Sivumäärä:	132

TIIVISTELMÄ: Euroopan komission perustama tekoälyä käsittelevä korkean tason asiantuntijaryhmä (AI HLEG) määrittelee luotettavan tekoälyn eettiset ohjeet. Näissä eettisissä ohjeissa kuvataan seitsemän keskeistä periaatetta, joita tekoälykehittäjien on noudatettava tekoälyn eettisen ja vastuullisen kehittämisen varmistamiseksi. Tekoälyn AI HLEG -asiantuntijaryhmän ohjeet on siirretty suoraan vuonna 2024 hyväksytyyn EU:n tekoälylakiin. Luotettava tekoäly viittaa tekoälyjärjestelmiin, jotka ovat selitettäviä, oikeudenmukaisia, tulkittavissa olevia, robusteja, läpinäkyviä, turvallisia ja suojattuja. Tämän opinnäytetyön keskeinen kysymys on, miten nämä seitsemän AI HLEG -asiantuntijaryhmän määrittelemää tekoälyn keskeistä eettistä periaatetta voidaan integroida terveydenhuollon innovaatio- ja tuotekehitystyöhön ja edelleen osaksi projektinjohtamisen prosesseja ja hyviä käytäntöjä. Tekoälyn tullessa yhä olennaisemmaksi osaksi sekä tuotekehitysprosesseja että niiden tuloksia, läpinäkyvän, jäsenneilyn ja eettisesti perustellun tekoälyhallinnan kysyntä kasvaa. Tämä opinnäytetyö vastaa tähän tarpeeseen tutkimalla niin kutsuttua Z-katselmuksen menetelmää (eng. Z-Inspection® method) ja sen soveltuvuutta tekoälyn luotettavuuden arviointiin selvittämällä, miten tällaiset arvioinnit voidaan integroida projektinjohtamisen elinkaareen. Tämän opinnäytetyön vaikutukset ulottuvat laajalle ja eri ammattiryhmiin. Käytännön ammattilaisten kannalta esitetyt tulokset korostavat tarvetta kehittää tekoälyn hallintaosaamista, edistää vastuullista tekoälykulttuuria sekä luoda toiminnallisia rakenteita tekoälykehitystyön sosioteknisiä arviointeja varten. Päättäjien kannalta tutkimus korostaa selkeämpien sääntelyohjeiden, tekoälyn vaatimusten yhdenmukaistamisen sekä organisaatioiden valmiuksien kehittämisen merkitystä terveydenhuollossa. Tutkijoille tämä opinnäytetyö tunnistaa uusia tutkimusmahdollisuuksia luotettavan tekoälyn käyttöönotossa osana projektinhallintaa sekä organisaatioiden kypsyiden arviointia tekoälyn hallinnassa. Kaiken kaikkiaan tämä opinnäytetyö osoittaa, että luotettavan tekoälyarvioinnin integrointi projektinhallintaan on sekä mahdollista että välttämätöntä. Tekoälyn jatkaessa terveydenhuollon innovaatio-työn muokkaamista, organisaatioiden on omaksuttava kontekstierkkiä johtamismenetelmiä turvallisuuden, läpinäkyvyyden ja yhteiskunnallisten arvojen mukaisuuden varmistamiseksi. Yhdistämällä eettiset periaatteet käytännön toteutukseen tämä työ edistää strategisia projektinhallintakäytäntöjä tukien luotettavaa tekoälyn innovointi- ja kehitystyötä. Ehdotettu Z-katselmuksen integrointi projektinhallinnan elinkaareen tarjoaa strukturoidun polun tekoälyn käyttöönotolle ja tukee samalla sen innovointia, luotettavuutta ja EU-lainsäädännön mukaisen sääntelyn noudattamista.

AVAINSANAT: Etiikka, innovaatiot, luotettavuus, projektinhallinta, sääntely, tekoäly ja terveydenhuolto.

To My Mother

Contents

Preface	12
1 Introduction	13
1.1 Background	13
1.2 Problem Statement	15
1.3 Research Questions and Hypotheses	16
1.4 Research Gap	17
1.5 Author's Contribution	18
1.6 Structure of the Thesis	18
2 Fundamentals of Trustworthy Artificial Intelligence	20
2.1 Trust and Trustworthiness	20
2.2 Trustworthy Assessment Model	22
2.3 On Trustworthy Assessment of Artificial Intelligence Systems	25
2.4 The Z-Inspection® Process	29
2.5 On Implementation of AI Regulatory Processes in Healthcare	30
3 Project Management under EU AI Regulation	33
3.1 Principles of EU AI Regulation	33
3.2 Current Role of Artificial Intelligence in Project Management	37
3.3 Artificial Intelligence Trends in Project Management	40
3.4 Impact of Artificial Intelligence on Project Management	43
3.5 Risks, Opportunities, and Trustworthiness Concerns	45
4 Machine Learning in Healthcare Innovation	47
4.1 Federated Machine Learning	47
4.2 On Trustworthy Explainable Artificial Intelligence	47
4.3 Foundations of Trustworthy Federated Machine Learning	49
4.3.1 Privacy and Security	50
4.3.2 Accountability, Bias, Fairness, and Transparency	51
4.3.3 Explainability and Interpretability	52
4.3.4 Technical Robustness	53

4.3.5	Data Integrity, Regulatory Compliance, and Trust	54
4.4	Technologies Supporting Trustworthy Federated Machine Learning	55
4.4.1	Internet of Medical Things	56
4.4.2	Blockchain Architecture	56
4.4.3	Trusted Execution Environment	57
5	Methods	58
5.1	Trustworthy AI Impact Assessment in the FEHLS Project	58
5.1.1	Study Design	58
5.1.2	Data Sources and Processing	59
5.1.3	Development of Federated Machine Learning Model	60
5.1.4	Design of Trustworthy Artificial Intelligence Assessment	60
5.1.5	Trustworthy Assessment Extensions	60
5.2	Design of the Proposed Project Management Framework	61
6	Results	65
6.1	Findings of Trustworthy AI Assessment in the FEHLS Project	65
6.1.1	Priority Actions	66
6.1.2	Identified Core Assessment Challenges and Recommendations	66
6.1.3	Strengths of Z-Inspection®	68
6.2	The Proposed Project Management Framework	69
6.2.1	Initiation Phase	69
6.2.2	Planning Phase	69
6.2.3	Execution Phase	70
6.2.4	Monitoring and Controlling Phase	70
6.2.5	Closing Phase	71
6.2.6	The Big Picture of the Proposed Framework	71
6.2.7	Strategic Benefits	73
6.3	Analytical Reflection on the Framework Supporting the Results	74
7	Discussion	78
7.1	Liability and Regulatory Aspects	78
7.2	Insights on Education to Support Artificial Intelligence Regulation	81

7.3	Geopolitical Competitive Context	83
7.4	Adopting Trustworthy AI in Healthcare Innovation Projects	85
7.4.1	Trustworthy Artificial Intelligence as a Practical Imperative	85
7.4.2	Socio-Technical Risks and Ethical Considerations	86
7.4.3	Integrating Trustworthiness Assessment into Project Management	86
7.4.4	The Role of the Z-Inspection® Method	87
7.4.5	Organizational Readiness and Future Outlook	87
7.5	Implications for Practice	88
7.6	Implications for Policy	90
7.7	Implications for Research	92
7.8	Study Limitations	93
8	Conclusion	96
	References	99
	Appendices	113
	Appendix 1: Identified Claims, Evidence and Reasonings in the FEHLS Study	113
	Appendix 2: Interview Questions and Responses	118
	Appendix 3: Trustworthy AI Assessment Template for Project Managers	128

Figures

Figure 1: Main components of trustworthy AI. In the chart, a model refers to a mathematical system that makes predictions or decisions. A process refers to the entire lifecycle of an AI system: how it is designed, developed, tested, deployed, and monitored. Additionally, data is the information used to train, validate, and operate the model. Adapted from Alzubaidi et al. (2023).	21
Figure 2: An overview of the trust development models, including TrAM's suggestions. Adapted from Schlicker et al. (2025).	24
Figure 3: The requirement list metaphor describes the relationship between individual standards, actual trustworthiness, and perceived trustworthiness. Adapted from Schlicker et al. (2025).....	25
Figure 4: The depicted framework by Wu & Liu (2023) categorizes new technologies based on their uncertainties into “art,” “craft,” and “science.” Greater technological uncertainties can bring great benefits, but also potential dangers. Adapted from Wu & Liu (2023).	34
Figure 5: Common underlying technologies of artificial intelligence in project management. Adapted from Vergara et al. (2025).	41
Figure 6: Benefits of artificial intelligence in project management, according to Shoushtari et al. (2024).	44
Figure 7: Framework for T-FML. Adapted from AISC (2024). Arrows with solid lines represent technology, data or information flows, while arrows with dashed lines represent control or management.	50
Figure 8: Diagram of the project phases using the trustworthy AI perspective proposed.	72
Figure 9: Regulatory challenges in artificial intelligence. Adapted from Scherer (2015).	80

Tables

Table 1: A summary of regulatory proposals from around the world, each adopting unique rationales and approaches to address the diverse concerns raised by AI, according to Wu & Liu (2023). 14

Table 2: Identified claims, evidence and reasonings (CERs) of the developed FML systems. The claims for the Norwegian and Danish systems were derived directly from information provided by the developers in the co-design studies. Adapted from Kringen et al. (2026). 113

Abbreviations

AI	Artificial Intelligence
AI HLEG	High-Level Expert Group on Artificial Intelligence
AI RMF	Artificial Intelligence Risk Management Framework
ALTAI	The Assessment List for Trustworthy AI
CER	Claims-Evidence-Reasoning
CFML	Centralized Federated Machine Learning
DFML	Distributed Federated Machine Learning
DL	Deep Learning
EHR	Electronic Health Record
EU	European Union
FedAvg	Federated Averaging
FEHLS	Federated Ethical Healthcare Learning Sandbox
FML	Federated Machine Learning
GDPR	General Data Protection Regulation
GenAI	Generative AI
GP	General Practice
HTTP	Hypertext Transfer Protocol
IEEE	Institute of Electrical and Electronics Engineers
IoMT	Internet of Medical Things
ISO	International Organization for Standardization
ITU	International Telecommunication Union
IVDR	In Vitro Diagnostic Medical Devices
MCAS	Maneuvering Characteristics Augmentation System
MDR	Medical Device Regulation
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MQTP	Message Queuing Telemetry Transport
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing

PM	Project Manager
SAIs	Supreme Audit Institutions
SHAP	SHapley Additive exPlanations
SME	Small or Medium-sized Enterprise
TEE	Trusted Execution Environments
T-FML	Trustworthy Federated Machine Learning
TrAM	Trustworthiness Assessment Model
TRL	Technology Readiness Level
T-XAI	Trustworthy Explainable Artificial Intelligence
UN	United Nations
WHO	World Health Organization
XAI	Explainable Artificial Intelligence

Preface

The practical work for this thesis was done at Tampere University in the summer of 2025. Four months of working in the Federated Ethical Healthcare Learning Sandbox (FEHLS) project introduced me to the Z-Inspection® method and the principles of federated machine learning, as well as to the excellent FEHLS project team in the Nordic countries. The FEHLS project was funded by the *Life Science and Health Tech* and *AI and Data* programs of the *Nordic Council of Ministers for Digitalisation*, which aims to promote cooperation between the Nordic countries, Estonia, Latvia and Lithuania.

My studies in industrial management at University of Vaasa consisted of strategic project management and economics. My major, strategic project management, naturally covers the main topic of this thesis. The influence of economics as my minor is clearly visible in the text and especially in the discussion section of this thesis. My previous education and work experience in digital signal and image processing, the engine of modern machine learning techniques, have given me an excellent foundation to understand modern artificial intelligence-based technologies, without forgetting the management insight brought by my previous managerial positions in university administration. This expertise helped me a lot in writing my thesis.

I would like to sincerely thank my esteemed supervisors Professor Mark van Gils from Tampere University and Degree Program Director, Dr. Marko Mäkilouko from University of Vaasa, for making this thesis possible.

I am also grateful to the members of the FEHLS project team and the members of the Decision Support for Health research group. Without them, I would not have had the opportunity to learn all this.

Finally, I would like to thank my good old friends Jukka, Kata and Toni for their support over the years. Their genuine support has had a huge impact on my goals to complete a second degree and advance my career.

Antti P. Happonen

D.Sc. (Tech.)

Tampere, the 20th of March 2026

1 Introduction

This thesis examines the integration of ethical principles of trustworthy artificial intelligence (AI) into healthcare product processes in innovation and development work. Building on the guidelines established by the High-Level Expert Group on Artificial Intelligence (AI HLEG) and their incorporation into the EU AI Act, the study focuses on seven key principles of trustworthy AI in project management. Trustworthy AI consists of explainability, fairness, interpretability, robustness, transparency, safety and security. The central objective is to explore how these principles can be embedded in project management practices, and to evaluate the applicability of the so-called Z-Inspection® method for assessing AI trustworthiness within innovation and development processes.

1.1 Background

A range of artificial intelligent regulatory proposals has emerged around the globe, each grounded in its own logic and strategy to confront the many challenges posed by AI. While policymakers around the world are trying to create regulatory principles based on the experiences of previous technologies, the unique complexity of AI makes this particularly difficult. Given how quickly AI is advancing and how many different dimensions it reaches, more structured and comprehensive regulatory approaches are becoming increasingly necessary, as Wu & Liu (2023) note.

It is worth considering whether legislators should aim to regulate AI as a unified whole or whether they should, instead, focus on regulating specific uses of AI in specific sectors or domains, as Finocchiaro (2024) reminds us. The European Union (EU) AI regulatory framework adopts the first option, which is a broad and multidisciplinary approach. Many international organizations, however, support the opposite strategy, arguing that it is more effective to regulate individual AI applications, or more precisely, their impact within the specific areas where they are used. Regulatory proposals regarding artificial intelligence and their justifications in different regions around the world are summarized in **Table 1**.

Table 1: A summary of regulatory proposals from around the world, each adopting unique rationales and approaches to address the diverse concerns raised by AI, according to Wu & Liu (2023).

Region	Rationale	Approach	Benefits	Challenges
USA	Industry-specific	Industry-specific	Preserves existing regulatory frameworks, leverages industry-specific expertise	Risk of inconsistency, overlap and absence
EU	Risk-based	Centralized	Offers a holistic view of AI regulation, sets clear and unified standards	Lack of adaptability and efficiency
UK	Context-specific	Contingent	Recognizes the importance of the application context, allows for flexibility in regulation	Lack of predictability, standardization, and transparency, potential for oversupply of regulatory discretion
China	State-controlled	State-controlled	Coordinate the rapid advancement of technology	Vulnerable to significant risks
International	Intelligence hub	Global standardization	Provides broad applicability	Can be slow to adapt to rapid advancements

The EU AI Act (Regulation (EU) 2024/1689, 2024) is European law governing artificial intelligence marking the first broad and universal AI regulation introduced by a major global authority. The goal is both to establish a unified regulatory model and to prevent fragmentation within the Union. This reflects the broader strategic vision of EU lawmakers, who ultimately seek to build a cohesive European digital market. As is often the case, the regulatory initiative serves geopolitical aims by attempting to extend the reach of EU

standards. The pursuit of European “*digital sovereignty*,” a concept that operates externally toward other global powers, and internally by influencing the policies of EU Member States, as Finocchiaro (2024) emphasizes.

According to the European Commission’s Explanatory Memorandum, “*It is in the Union’s interest to preserve the EU’s technological leadership.*”, as Finocchiaro (2024) reports. However, it must be acknowledged that the EU does not have a global leadership position in the field of artificial intelligence, nor is it one of the world's largest producers of artificial intelligence technologies. Within this geopolitical landscape, the EU’s strategy is instead to position itself as a global leader in regulatory governance, aiming to make the European regulatory model a worldwide benchmark. This dynamic often described as the “*Brussels effect*” (Finocchiaro, 2024). The objective is not to compete with the United States and China in technological output, but rather in shaping the rules that govern AI.

Practically, the EU AI regulation categorizes AI systems into three levels of risk (EU AI Act, 2026). The first risk covers AI use cases deemed unacceptable, such as government-operated social scoring systems like those in China, which are totally prohibited in the EU. The second category includes high-risk systems, like software that screens and ranks job applicants, which must meet strict legal standards. By default, many AI-supported healthcare systems are included in this high-risk category. The third category consists of AI applications that are neither banned nor classified as high-risk, which face minimal or no regulation, as the EU AI Act (2026) clarifies.

1.2 Problem Statement

Since many AI-driven healthcare tools will automatically be classified as high-risk, it means that product innovation and development teams in healthcare-related projects need to (EU AI Act, 2026):

- Expect third-party conformity assessments, not only self-assessment.
- Plan for robust risk or quality management systems, documentation, and technical files.
- Prepare for post-market monitoring and incident reporting.

That is, all these key aspects must be considered in project work for AI-driven healthcare products. In this thesis, we stress this very topic: trustworthy AI in project management.

1.3 Research Questions and Hypotheses

The practical component of this thesis was completed at Tampere University during the summer of 2025. This four-month work package was a part of the Federated Ethical Healthcare Learning Sandbox (FEHLS) project, funded by *Nordic Innovation* through the Nordic Applied Ethical AI Consortium. The FEHLS project showcases how federated machine learning (FML) can be used with unstructured clinical data across the Nordic countries and organizations in a way that aligns with trustworthy AI principles (Nordic Innovation, 2026).

As a basis of this thesis, developed in the FEHLS project, we present a collaborative approach to evaluating the trustworthy FML system in both Norway and Denmark. Our evaluation is based on the Z-Inspection[®] framework, a method designed to assess the trustworthiness of AI systems in real-world contexts (Zicari et al., 2021). To address differences in data environments, systems design, and clinical practices, we conducted separate co-design activities, involving developers, clinicians, and Z-Inspection[®] experts for the Norwegian and Danish development teams.

This thesis is guided by the following three **Research Questions (RQs)**:

RQ1: What specific benefits does the Z-Inspection[®] methodology offer for the systematic assessment of trustworthy artificial intelligence systems?

RQ2: In what ways can the Z-Inspection[®] methodology be effectively integrated into established project management processes to enhance the governance of AI-related innovation projects?

RQ3: What are the practical, regulatory, and organizational implications of integrating the Z-Inspection[®] methodology into project management processes within the context of current EU legislation on artificial intelligence?

Our **Research Hypotheses (RHs)** are as follows:

RH1 (related to **RQ1**): The application of the Z-Inspection[®] methodology provides measurable improvements in the robustness, transparency, and ethical compli-

ance of trustworthy AI assessments compared to conventional assessment approaches.

RH2 (related to **RQ2**): Integrating the Z-Inspection® methodology into project management processes enhances the effectiveness of AI project governance by introducing structured decision points, clearer accountability, and more comprehensive risk-mitigation practices.

RH3 (related to **RQ3**): Alignment of the Z-Inspection® methodology with EU legislative requirements facilitates greater regulatory compliance and reduces uncertainty regarding legal liability and organizational responsibility in AI-driven projects.

While researching, we discovered that in addition to the three **Research Questions**, additional hypotheses could streamline our approach. The additional **Research Hypothesis** are as follows:

RH4: Organizations adopting a Z-Inspection®-integrated project management framework will demonstrate higher levels of organizational readiness for trustworthy AI implementation, particularly in socio-technical and ethical dimensions.

RH5: The integration of the Z-Inspection® methodology into project management practices supports better stakeholder understanding and acceptance of trustworthy AI requirements, especially in regulated sectors such as healthcare.

1.4 Research Gap

To the best of our knowledge, the integration framework proposed in this thesis addresses an existing *research gap* and offers a novel analytical perspective through which the research questions are examined.

The initial academic results of the FEHLS project were published in two abstracts (Happonen and van Gils, 2025; Happonen et al., 2025) and in an article by Kringen et al. (2026). Our research in this thesis further integrates the results of the FEHLS project into a traditional project management framework, as described, for example, by the Project Management Institute (PMI, Chapter 2, 2021). Our framework is targeted at AI-driven healthcare product innovation and development projects in accordance with EU regula-

tions. This integration framework is achieved through a literature review and supported by an analysis of prompt engineering interview responses.

1.5 Author's Contribution

The author participated in a four-month work package for the FEHLS project (FEHLS, 2026) at Tampere University in the summer of 2025. The work included, among other things, a literature review, participation in the Z-Inspection® assessment team and preparation of the first draft of the article by Kringen et al. (2026) before its final analysis and conclusions. The FEHLS project framework from the Z-Inspection® perspective was presented in Espoo (Happonen et al., 2025), where the author drafted the abstract.

During the author's four months of work experience with the Z-Inspection® method, the idea of integrating the method into project management processes to promote innovation and development in accordance with the EU artificial intelligence regulation emerged in the summer of 2025. At that time, the author was considering a possible thesis topic on strategic project management, which is his major in industrial management at University of Vaasa. The first thoughts on this topic were published in the abstract of a seminar at Tampere University (Happonen & van Gils, 2025).

This thesis presents a framework for integrating Z-Inspection® and project management, demonstrating the author's own academic contribution. To our knowledge, there is no or little previous academic research on the topic, as indicated by our extensive literature review. However, our results illustrate the general steps of the framework, and it opens more avenues of research than it provides answers. The formulation of the proposed framework into a coherent and practical model with detailed practical findings is beyond the scope of this thesis.

1.6 Structure of the Thesis

In this thesis, **Chapter 2** outlines the foundational concepts related to trustworthiness in artificial intelligence. **Chapter 3** examines the principles of AI within the context of project management under the EU regulatory framework, drawing on our extensive review

of the relevant literature. **Chapter 4** presents the findings of our literature survey on federated machine learning and its role in healthcare innovation.

Chapter 5 details the research methodology employed in this thesis. Subsequently, **Chapter 6** introduces our proposed project management framework incorporating trustworthy AI processes, and more importantly, the results obtained in this thesis. Within **Chapter 7**, we discuss the diverse perspectives that inform the development of the framework. Specific attention is given to issues, such as liability and regulation, educational considerations necessary to support AI governance, and the geopolitical dynamics influencing AI adoption in healthcare innovation (Christensen et al., 2008). Moreover, In **Chapter 7**, we address socio-technical risks and ethical dimensions relevant to organizational readiness in terms of this thesis topic. Our discussion in **Chapter 7** includes also the practical, policy-related, and research-oriented implications for our findings, in addition to other issues we found relevant for this study. Furthermore, the limitations of this thesis are articulated prior to the final conclusions in **Chapter 8**.

Appendix 1 provides a summary of selected Z-Inspection® results from the FEHLS project. **Appendix 2** contains the interview questions alongside chatbot-generated simulated responses. **Appendix 3** presents a template for trustworthy AI assessment intended for practical use in project management.

2 Fundamentals of Trustworthy Artificial Intelligence

2.1 Trust and Trustworthiness

The rapid expansion of artificial intelligence (AI) use in recent years has brought the issue of trustworthiness to the forefront. Policymakers and researchers are increasingly emphasizing concerns related to trust, transparency, and ethical considerations in AI-driven healthcare systems. Numerous studies indicate that when healthcare professionals and decision-makers trust AI technologies, the quality of medical services and patient care tends to improve, as Albahri et al. (2023) note.

In the medical field, patients place significant trust in their physicians and care providers, often in situations where decisions may influence life, death, or long-term disability. Linguistically, trust as a term reflects the attitude that individuals hold toward other people, tools, or systems, such as AI-enabled medical devices or software that they assume to be trustworthy. It is important to understand that, unlike trust, trustworthiness describes an inherent characteristic and not an attitude. This distinction between trust and trustworthiness (or reliability) highlights that a trustworthy individual or technical system is not necessarily trusted, and someone or something that is trustworthy can trust others who are not. Thus, trust and trustworthiness are related but fundamentally different concepts. Ideally, what we choose to trust should indeed be trustworthy, and truly trustworthy entities should earn our trust, as Hasani et al. (2022) describe.

When discussing trustworthy AI (see **Figure 1**), several important questions may arise in our minds. For example, in their review Kaur et al. (2022) pose the following questions to support this thought process:

What conditions are necessary for artificial intelligence to be considered trustworthy?

Which guidelines and policy frameworks are needed to effectively govern AI systems?

Why does human involvement remain crucial in an era increasingly shaped by AI?

Which factors determine whether AI-driven decisions are viewed as legitimate and acceptable?

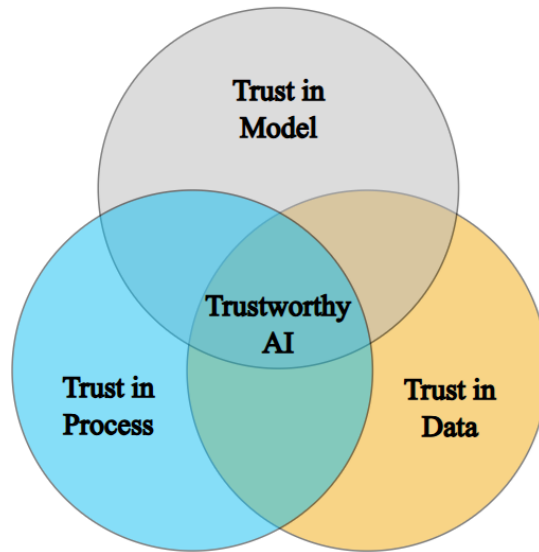


Figure 1: Main components of trustworthy AI. In the chart, a model refers to a mathematical system that makes predictions or decisions. A process refers to the entire lifecycle of an AI system: how it is designed, developed, tested, deployed, and monitored. Additionally, data is the information used to train, validate, and operate the model. Adapted from Alzubaidi et al. (2023).

Because many concepts in the trustworthy AI field can be interpreted in distinct ways, it is essential to clarify and define the terminology. In this section, we outline first two key terms that are relevant in AI-related studies. Kaur et al. (2022) define them:

Artificial Intelligence: AI refers to the development of machines capable of performing tasks that require intelligence. AI can be described as the imitation of human behavior and decision-making abilities. In essence, AI involves training machines to carry out functions that typically rely on human cognitive skills.

Black-Box Problem: The black-box problem arises when a system's internal processes are opaque, making it difficult to understand its structure, functioning, or implementation. As AI systems grow increasingly complex, they become harder to interpret. This lack of transparency undermines trust, as it becomes challenging to justify or explain the system's outputs.

Alternatively, using legal terminology, rather than treating "*artificial intelligence systems*" as a technical category, the term can simply refer to the relationship between two elements as (Scherer, 2015):

A software model that continually updates its predictions as it receives new information and generates output.

The link between that iterative process and the actions carried out either by another software component or by an external actor.

We will clarify other trustworthy AI-related terms in later sections of this thesis.

2.2 Trustworthy Assessment Model

In the context of AI, trust develops through a person's own evaluation of how dependable a system actually is. This evaluation is known as the trustworthiness assessment process. When this assessment is accurate, individuals or users can form realistic expectations about what the technical system can and cannot do. This reasoning allows them to make well-informed decisions about how much to rely on the system, as Schlicker et al. (2025) describe.

If trust is overestimated, users may become less attentive and lose an appropriate level of oversight. If trust is underestimated, they may disregard correct or valuable system outputs. These challenges connect directly to research on "*calibrated trust*," which, as described by Schlicker et al. (2025), aims to optimize how humans and AI systems work together.

Trusting a person or a system involves three key elements, according to Hasani et al. (2022). The elements can be listed as:

1. Accepting a degree of vulnerability because one's own interests depend on the trustee's actions.
2. Believing that the person or technology has the capability to achieve the intended objective.
3. Assuming they are willing to carry out the task as expected.

Schlicker et al. (2025) explain how their Trustworthiness Assessment Model (TrAM) builds on and extends existing trust frameworks. They clarify in their article how the

central concepts involved in trust formation relate to one another and differentiate the TrAM terminology from similar notions used inconsistently across prior academic studies. The aim of Schlicker et al. (2025) is to improve conceptual precision in trust research, addressing ongoing concerns about the inconsistent measurement of trust-related constructs and broader critiques of trust as a scientific topic.

Schlicker et al. (2025) draw clear distinctions between actual versus perceived trustworthiness, trust propensity, trust itself, and observable trust behavior. They also define, and refine, the core components of trustworthiness assessment, i.e., actual trustworthiness, perceived trustworthiness, system attributes, and individual evaluative standards, while highlighting the importance of both micro- and macro-level influences on the trustworthiness assessment process.

At the micro level, TrAM draws on psychological theories describing how people reason and how they judge the characteristics of others. Prior research, as noted by Schlicker et al. (2025), indicates that certification bodies and expert communities significantly shape public trust in AI systems. Consequently, these actors may also influence how individuals form trust in AI-based technologies, underscoring the role of social cues and heuristics in evaluating an AI system's trustworthiness.

To capture how the judgments of various stakeholders influence one another, Schlicker et al. (2025) introduced a macro-level perspective. They conceptualize system-level trustworthiness as emerging from a network of micro-level assessments made by different individuals or trustors, each contributing new signals that shape the perceptions of others.

Finally, Schlicker et al. (2025) outline practical implications of TrAM for system designers, human trustees, and policymakers. They emphasize the model's relevance considering ongoing regulatory developments in AI standardization and certification, including initiatives such as the European Artificial Intelligence Act based on the proposals of the High-Level Expert Group on Artificial Intelligence Group (AI HLEG, 2020), as well as other similar regulatory efforts around the world.

Figure 2 illustrates the trust development model with two main components, as Schlicker et al. (2025) present:

1. The progression from a trustor's perceived trustworthiness of a system to their actual trusting behavior, following the Organization's Trust Model (shown on the right side of **Figure 2** in light grey); and
2. the trustworthiness assessment process, through which trustors form their perceptions of trustworthiness based on the system's actual trustworthiness (depicted on the left side of **Figure 2**).

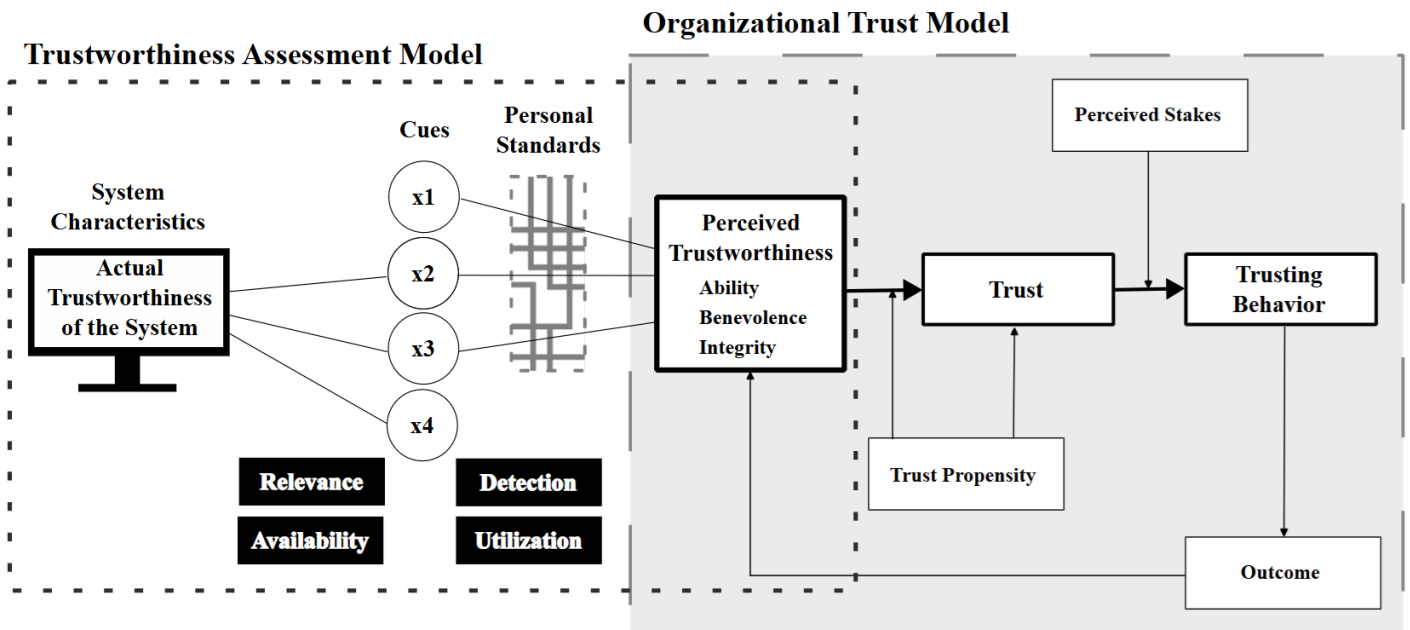


Figure 2: An overview of the trust development models, including TrAM's suggestions.

Adapted from Schlicker et al. (2025).

Schlicker et al. (2025) also highlight a key aspect of how trustworthiness information spreads: the distinction between primary and secondary cues. Primary cues originate directly from the system itself, whereas secondary cues reflect another stakeholder's evaluation of the system's trustworthiness. These secondary cues transmit information that is shaped or "*colored*" by the individual standards and interpretations of the stakeholders who produced them.

Schlicker et al. (2025) emphasize that what counts as a trustworthy system differs from one trustor to another, as each person relies on their own standards for evaluating trustworthiness. These individual standards essentially answer the question: "*What makes a*

system trustworthy for me?" A useful way to think about these standards is as a personalized checklist that contains all the elements a trustor considers necessary for a system to be fully trustworthy, as illustrated in **Figure 3**.

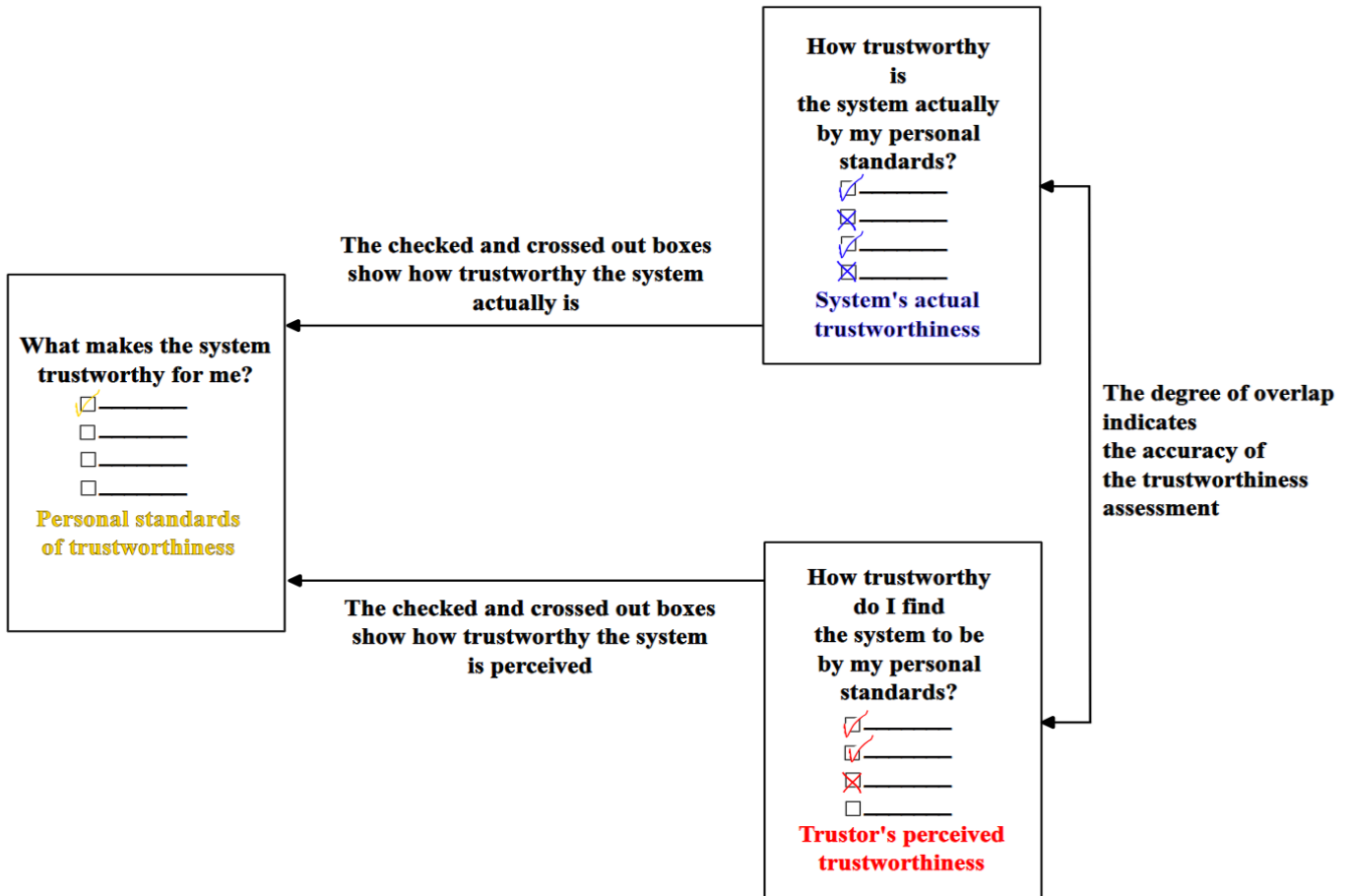


Figure 3: The requirement list metaphor describes the relationship between individual standards, actual trustworthiness, and perceived trustworthiness. Adapted from Schlicker et al. (2025).

2.3 On Trustworthy Assessment of Artificial Intelligence Systems

It can be argued that there is still significant uncertainty regarding the application of risk-based methods and ethical principles in real-world AI contexts, as Moraes (2025) notes. This lack of clarity has drawn also justified criticism, for instance, from civil society groups. The criticism has contributed to the emergence of new regulatory tools for managing disruptive technologies, most notably the so-called regulatory sandboxes, which relate

to responsible innovation, or in other words, the *Collingridge dilemma* (Genus & Stirling, 2018). The European Commission's AI Act (Regulation (EU) 2024/1689, 2024) explicitly authorizes competent authorities to establish such controlled testing environments or regulatory sandboxes, as Moraes (2025) points out.

Kemmerzell et al. (2025) conducted a comprehensive review of existing approaches designed to support the systematic implementation of trustworthiness in AI systems. By analyzing 18 major frameworks developed by national authorities, international bodies, and research institutions, they sought to clarify how AI trustworthiness is currently conceptualized and assessed. Stahl et al. (2023) have systematically reviewed up to 38 AI impact assessment methods (which are beyond the scope of this thesis due to the large number of methods).

Despite the approaches' differences, these trustworthiness assessment frameworks share several notable similarities. Most prominently, they adopt a risk-based perspective that emphasizes governance and risk management. They also converge on a set of recurring themes, including fairness, transparency, trust, security, cybersecurity, and accountability. Privacy and data protection appear consistently as well, sometimes framed as a requirement (EU AI Act, 2026), sometimes as a principle (OECD, 2025), and sometimes as a system feature (Tabassi et al., 2023). These very elements are common across many AI trustworthiness frameworks, as Moraes (2025) notes.

Kemmerzell et al. (2025) categorize existing AI trustworthiness assessment frameworks into three groups based on their purpose and maturity:

1. Guidelines and ethical principles.
2. Roadmaps, standards, and regulations.
3. Checklists, catalogues, and evaluation tools.

Ethical frameworks, such as the EU's ALTAI guidelines by the AI HLEG group (ALTAI, 2020), offer processes for navigating conflicts between ethical principles in practice, but they often lack concrete, context-specific guidance. As Bærøe et al. (2020) point out, these frameworks also do not always address tensions between their own ethical aims (e.g., promoting justifiable AI) and other societal goals, such as economic development or progress toward the *Sustainable Development Goals* by the United Nations (UN, 2023).

Buruk et al. (2020) compare the EU's AI HLEG guidelines with other influential documents that reflect different institutional perspectives on AI. For example, they find out that the *Montreal Declaration for Responsible AI* stems from academic discourse (Montreal Declaration, 2018), while the *Asilomar AI Principles* from the Future of Life Institute reflect a blend of private-sector and research viewpoints (Asilomar AI, 2017).

Hickman and Petrin (2021), on the other hand, explore how guidelines like EU regulations may influence corporate law and governance. They argue that further analysis is needed to determine how well these regulatory proposals align with existing corporate governance principles. Nonetheless, they maintain that such guidelines provide a valuable foundation for companies seeking to develop trustworthy AI, at least until more binding regulatory instruments are introduced.

The following list briefly presents the generally applied trustworthy AI assessment frameworks identified by Kemmerzell et al. (2025).

IEEE CertifAIED: The Institute of Electrical and Electronics Engineers' (IEEE') CertifAIED is a certification framework. It demonstrates that a professional has the competence and practical experience needed to apply IEEE's ethical AI frameworks and associated methodologies in their work (IEEE, 2021).

The Fraunhofer IAIS AI Test Catalog: To mitigate risks and strengthen public trust in artificial intelligence, several bodies, such as the EU's High-Level Expert Group on AI (AI HLEG, 2020) and the German Federal Government's Data Ethics Commission, have issued broad guidelines for AI development. These guidelines, however, often remain high-level and lack concrete, actionable requirements for developers and organizations. The *Fraunhofer IAIS AI Test Catalog* directly addresses this gap by offering a structured framework that translates abstract quality principles into application-specific testing criteria (Poretschkin et al., 2023).

capAI: capAI provides a structured methodology for ensuring and demonstrating that AI systems align with an organization's ethical values. It adopts a lifecycle perspective, evaluating practices across five stages: design, development, evaluation, operation, and decommissioning (Floridi et al., 2022).

The capAI procedure consists of three components:

- An internal assessment protocol,
- a summary data sheet, and
- an external scorecard.

Together, these elements help both developers and users assess ethical considerations at each stage of the AI lifecycle and verify compliance with the core requirements outlined in the AI Act.

The Assessment List for Trustworthy AI (ALTAI): The ALTAI guidelines by the EU's AI HLEG group translates high-level AI ethics principles into a practical, dynamic checklist that guides developers and adopters in implementing trustworthy AI. It provides concrete steps for self-assessment, helping organizations ensure that users benefit from AI systems without being exposed to unnecessary risks (ALTAI, 2020).

The Kortum's Procedural Model: Kortum et al. (2022) propose a procedural model that introduces users to a comprehensive set of measures aimed at preventing discrimination, insufficient accuracy, inadequate data protection, and lack of transparency in AI development. The model offers a foundational structure that can be expanded with more detailed analyses of individual steps and methods, thereby contributing to greater trust in AI systems.

Artificial Intelligence Risk Management Framework (AI RMF): The AI RMF 1.0, developed by the *U.S. National Institute of Standards and Technology (NIST)*, provides organizations with a flexible resource for managing AI-related risks and promoting responsible, trustworthy AI. The framework is voluntary, rights-respecting, sector-agnostic, and adaptable to organizations of all sizes and contexts (Tabassi, 2023).

The Newman's AI Trustworthiness Taxonomy: Building on the NIST AI RMF 1.0, Newman (2023) introduces an AI trustworthiness taxonomy designed to complement and support the framework. It includes 150 trustworthiness attributes, each derived from NIST's core trustworthiness dimensions:

1. Validity and reliability.
2. Safety, security, and resilience.
3. Accountability and transparency.
4. Explainability and interpretability.

5. Privacy preservation.
6. Fairness with mitigation of harmful bias.

Supreme Audit Institutions: In 2017, the Supreme Audit Institutions (SAIs) of Brazil, Finland, Germany, the Netherlands, Norway, and the United Kingdom signed a Memorandum of Understanding on data analytics. Recognizing that digitalization is reshaping government operations, the SAIs committed to collaborating on new methods and practices for effective auditing. At their 2019 conference in Finland, they agreed to jointly develop a document for auditing AI applications (Audit Algorithms, 2020).

VDE: The VDE SPEC provides a framework for describing socio-technical properties of systems that incorporate AI technologies. It is particularly relevant for products requiring a high degree of trustworthiness (Peylo et al., 2023).

2.4 The Z-Inspection® Process

In this thesis, we apply the Z-Inspection® methodology as a trustworthy AI assessment framework (<https://z-inspection.org/>) because the Z-Inspection® method offers a practical process for evaluating the trustworthiness of AI systems at any point in their lifecycle (Zicari et al., 2021). Furthermore, since the Z-Inspection® approach is originally grounded in the ALTAI (2020) guidelines, the methodology also aligns with the requirements of the current EU AI Act (Regulation (EU) 2024/1689, 2024).

In addition to the EU AI trustworthy criteria, Z-Inspection® also considers two other aspects (Zicari et al., 2022):

- Whether the AI ecosystem respects the values of Western modern democracy.
- Whether it avoids undue concentration of power.

Thus, for our purposes, the Z-Inspection® methodology provides a comprehensive socio-technical assessment method that evaluates ethical issues in specific use cases through scenario-based analysis. It can be used for co-design, self-assessment, or independent audits involving diverse stakeholder groups, as Zicari et al. (2021) describe.

The Z-Inspection® process consists of three phases (Vetter et al., 2023):

Set-up phase: verifying preconditions such as legal acceptability and absence of conflicts of interest, assembling a multidisciplinary expert team, and defining the scope and context of the assessment.

Assessment phase: iteratively developing socio-technical scenarios, identifying ethical tensions, validating claims with evidence, and mapping findings to the EU Trustworthy AI Framework using an “*open to closed vocabulary*” consensus approach.

Resolution phase: addressing identified ethical tensions, evaluating risks, proposing mitigation measures, and formulating recommendations for key stakeholders.

Z-Inspection® can be applied across the entire AI-system lifecycle: design, development, deployment, monitoring, and decommissioning (Zicari et al., 2022).

Worth noticing is that since 2021, FUTURE-AI has closely mirrored the Z-Inspection® approach, especially in the healthcare context. The FUTURE-AI framework for trustworthy AI in healthcare builds around six FUTURE principles: Fairness, Universality, Traceability, Usability, Reliability, and Explainability. The framework consists of an international, multi-stakeholder initiative and consensus guidance aimed at ensuring the development of trustworthy and deployable artificial intelligence in healthcare (Lekadir et al., 2025). In this thesis, we now focus on the implementation processes related to assessing the trustworthiness of artificial intelligence in healthcare innovations. In the practical part of the thesis, we evaluate the developed federated machine learning system by taking into account, among other things, ethical, environmental and socio-technical issues to ensure the trustworthiness of AI in innovation and development work in accordance with the EU AI Act.

2.5 On Implementation of AI Regulatory Processes in Healthcare

Health privacy regulations give individuals the authority to manage their personal medical data and set boundaries on who is allowed to access or receive it. In general, the ethical issues and challenges highlighted in reports such as by AI HLEG (2020) are shaping how AI will be developed and deployed in Europe and globally.

Since these principles are grounded in fundamental human rights, their relevance extends universally. The principles are closely aligned with the rights and obligations underpinning the UN Sustainable Development Goals (UN, 2023), which also inform strategies for evaluating trustworthiness in healthcare (Bærøe et al., 2020).

Radclyffe et al. (2023) call for clear and detailed regulations for assessing trustworthy AI and believe that such regulations are inevitable. Organizations such as *Institute of Electrical and Electronics Engineers (IEEE)*, *International Telecommunication Union (ITU)*, and *World Health Organization (WHO)* have initiated global standardization programs for the use of AI in healthcare. These initiatives establish standardized benchmarking frameworks that allow machine learning and AI models to be assessed in a uniform and transparent way, such as Albahri et al. (2023) address.

It goes without saying that industries like healthcare will benefit from adopting trustworthy AI practices as early as possible, gaining hands-on experience and positioning themselves as leaders in the responsible use of AI. Early implementation can foster greater trust and help avoid future conflicts among stakeholders. The focus in AI ethics has shifted from simply outlining high-level principles to ensuring these principles are put into action through concrete procedures and real-world practice, as Radclyffe et al. (2023) reminds us.

As noted, the EU regulation on trustworthy AI has been in force in the EU since 2024 (Regulation (EU) 2024/1689, 2024). Tools based on the EU AI Act framework play a significant role in reducing risks for organizations. The seven key elements of trustworthy AI required for the development and deployment of AI-powered tools can be summarized as follows (AI HLEG, 2020):

- 1) Human agency and oversight
 - I. Fundamental rights
 - II. Human agency
 - III. Human oversight
- 2) Technical robustness and safety
 - I. Resilience to attack and security
 - II. Fallback plan and general safety
 - III. Accuracy
 - IV. Reliability and reproducibility

- 3) Privacy and data governance
 - I. Privacy and data protection
 - II. Quality and integrity of data
 - III. Access to data
- 4) Transparency
 - I. Traceability
 - II. Explainability
 - III. Communication
- 5) Diversity, non-discrimination, and fairness
 - I. Avoidance of unfair bias
 - II. Accessibility and universal design
 - III. Stakeholder participation
- 6) Societal and environmental well-being, and
 - I. Sustainable and environmentally friendly AI
 - II. Social impact
 - III. Society and democracy
- 7) Accountability
 - I. Auditability
 - II. Minimizing and reporting negative impacts
 - III. Trade-offs redress

All seven rules listed are equally important and complementary and should be implemented and reviewed throughout the life cycle of an AI-enabled system. It is worth remembering that trustworthy AI is still a relatively new area of research, but it has gained attention in recent years as AI-driven tools have become widespread in healthcare. It is expected that the amount of research on the requirements for trustworthy AI in the healthcare literature will continue to grow in the future, as Albahri et al. (2023) forecast. The Z-Inspection® method used in our research follows these seven ethical principles.

3 Project Management under EU AI Regulation

The EU AI regulation entered into force in 2024, and most AI high-risk requirements become fully applicable around August 2026, with some transition periods continuing thereafter (MedTech Europe, 2025).

The EU AI Act introduces a structured, risk-based system for overseeing AI, placing the greatest regulatory burden on impactful applications, particularly in healthcare, where safety, reliability and accountability are crucial. Therefore, especially in the healthcare sector, the principles of the EU regulation must be considered in the project management workflow accordingly.

3.1 Principles of EU AI Regulation

The EU AI Act is the world's first comprehensive legal framework on AI, designed to ensure that AI systems are safe, transparent and respect fundamental rights. Its core idea is a risk-based approach, with stricter rules applied to systems with a higher potential impact (EU AI Act: Explorer, 2026). According to the EU AI Act, high-risk AI systems literally include those that (Regulation (EU) 2024/1689, 2024):

- Serve as a safety component or form part of a product regulated under the EU laws listed in the EU AI Act's Annex I (The EU AI Act: Annex I, 2026) and are required to undergo an independent conformity assessment under those same laws; *or*
- Fall under the use cases listed in the EU AI Act's Annex III (The EU AI Act: Annex III, 2026), *unless*:
 - The system only carries out a limited, procedural function.
 - It merely enhances the outcome of an already completed human task.
 - It identifies patterns or inconsistencies in decision-making but is not intended to replace or influence human judgment without appropriate human oversight; *or*
 - It performs a preliminary task that supports an assessment relevant to an Annex III use case.

- AI systems listed in Annex III are *automatically* considered high-risk if they involve profiling, meaning automated processing of personal data to evaluate aspects of a person’s life, such as job performance, finances, *health*, preferences, behavior, or movements.
- If a provider’s AI system falls under an Annex III category but the provider believes the system should not be classified as high-risk, they must document this justification before the system is launched or put into use.

As illustrated in **Figure 4**, a general technology regulatory framework provides a useful starting point for understanding the interplay between technological uncertainty and regulatory adaptability, as presented by Wu & Liu (2023). Each level of technological uncertainty requires an appropriate regulatory approach to guide development and mitigate the potential risks of that technology.

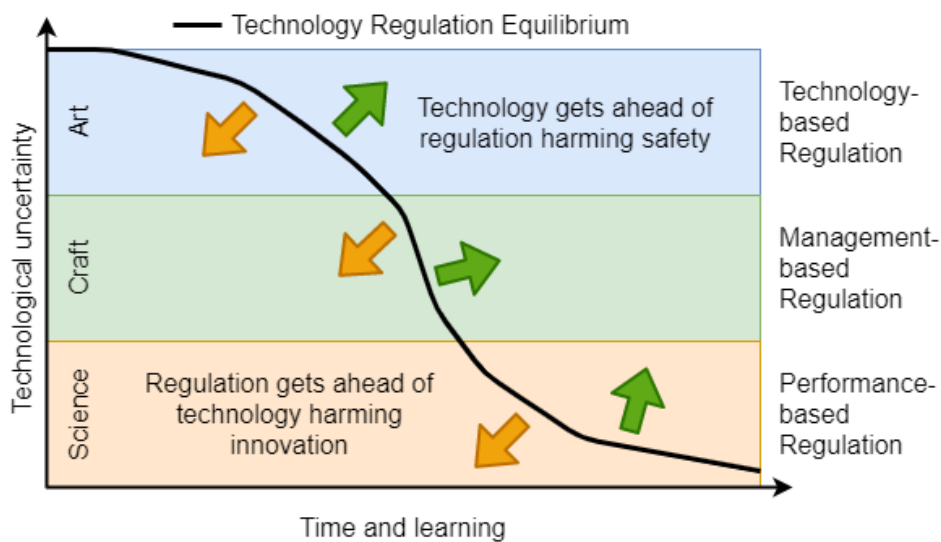


Figure 4: The depicted framework by Wu & Liu (2023) categorizes new technologies based on their uncertainties into “*art*,” “*craft*,” and “*science*.” Greater technological uncertainties can bring great benefits, but also potential dangers. Adapted from Wu & Liu (2023).

The next list covers the eight keys of how the EU AI Act (2026) should be considered in healthcare product development and project management.

- 1) Assume Most Healthcare AI Will Be High-Risk:** Because medical devices and healthcare-related software generally fall under existing EU product-safety frameworks like the *Medical Device Regulation* or MDR (Regulation (EU) 2017/745, 2017) and *In Vitro Diagnostic Medical Devices* or IVDR (Regulation (EU) 2017/746, 2017), many AI-driven healthcare tools will automatically be classified as high-risk.

This means that AI-driven development project teams need to expect third-party conformity assessments, not just self-assessment. The teams should plan for robust quality management systems, documentation, and technical files. Furthermore, they need to prepare for post-market monitoring and incident reporting.

- 2) Integrate AI Act Requirements Early in the Development Cycle:** The EU AI Act functions similarly to MDR: compliance cannot be added at the end. Healthcare projects should apply “*by design*” principles (e.g., transparency, robustness, data governance) from the start. The projects must include regulatory checkpoints in early-stage design reviews. Besides, the project teams need to allocate budget and time for conformity assessment steps.

- 3) Strengthening Data Governance and Training Data Quality:** The EU AI Act emphasizes requirements such as high-quality, representative, and bias-checked training data. This means clear documentation on how data is collected, labeled, and curated. Furthermore, traceability of datasets and preprocessing steps should be clearly described. For healthcare AI, this also intersects with GDPR (2016), clinical data standards, and ethical review processes.

- 4) Maintain Human Oversight Requirements:** Healthcare AI often supports clinical decisions, but it must not replace them without safeguards. Development project teams need to define clear human-in-the-loop roles. Also, they need to design interfaces that support explainability and clinician understanding. Furthermore, the project teams should document how human oversight prevents misuse or overreliance.

- 5) Prepare for Transparency Obligations:** High-risk healthcare AI will require clear information for users (clinicians, patients, regulators, etc.) about capabilities and

limitations. Also, logging and traceability of system behavior are mandatory. Clear descriptions of intended use and performance metrics should be described.

- 6) Address Profiling Concerns If Patient Data Is Analyzed:** If the AI profiles individuals to assess health conditions, behavior patterns, or lifestyle-related risks, it strengthens the classification as high-risk under Annex III rules. Developers must plan for strong safeguards against discriminatory outcomes. They need to make explicit documentation of profiling mechanisms. Justifications for why profiling is necessary and safe are also required.
- 7) Support Clinical Validation and Real-World Evidence:** The EU AI Act expects performance accuracy, robustness, and cybersecurity to be validated. In healthcare clinical validation studies must demonstrate safety and effectiveness. Besides, AI performance must be checked on diverse, real-world patient groups. Also, continuous performance monitoring after deployment becomes essential.
- 8) Ensure Alignment With MDR/IVDR:** The EU AI Act does not replace medical device law, but it layers on top of it. Developers should align: 1) risk management files, 2) clinical evaluation plans, and 3) post-market surveillance processes to satisfy both MDR/IVDR and the EU AI Act.

A practical summary for healthcare product teams can be listed as follows (EU AI Act, 2026):

- Treat the AI Act as a core regulatory requirement, not an add-on.
- Plan for high-risk classification and third-party evaluation.
- Build strong, transparent data and model documentation.
- Include explainability, human oversight, and monitoring in system design.
- Align processes with existing medical device regulations.

If properly implemented, the EU AI Act can improve trust, safety and the adoption of AI in healthcare, while reducing future regulatory friction. However, as the European trade association for the medical technology industry (MedTech Europe, 2025) points out, while the AI Act strengthens safety and accountability, insufficient alignment with existing medical regulations could lead to duplication of work, slower innovation, and delayed patient data, if clearer exceptions and harmonization measures are not implemented.

3.2 Current Role of Artificial Intelligence in Project Management

In project management workflow (Pica, Chapter 3, 2015), the five well-known phases or process groups consist of the following actions:

Initiate: Define a new project or phase by obtaining authorization and clarifying objectives, scope, and stakeholders.

Plan: Define scope, objectives, and an action plan. This includes schedules, budgets, resources, risk plans, quality plans, and communication plans.

Execute: Involve completing the work defined in the plan to achieve the project objectives: coordinating teams, managing resources, and ensuring deliverables are produced.

Monitor and Control: Monitor, review, and control project progress and performance. Ensures alignment with plan, manages change, and mitigates risks.

Close: Finalize all project activities, validates deliverables, completes documentation, and formally closes the project or phase.

In other words, project management is a complex field that requires careful planning, execution, and oversight to ensure projects succeed. In the past, project managers primarily depended on their own expertise, intuition, and analytical abilities to handle the challenges involved (PMI, Chapter 2, 2021). Today, however, the rise of artificial intelligence is transforming how projects are managed. AI provides powerful computational capabilities that can process large volumes of data, identify trends, and offer predictive insights that improve project efficiency and execution.

Generally, AI-driven tools enable project managers (PMs) to reduce cost overruns, anticipate potential risks, and optimize workflows. Research has shown that AI-enhanced project management systems strengthen real-time decision-making with sophisticated algorithms (Niederman, 2021).

The study by Jannat et al. (2024) indicate that the use of AI technologies is steadily increasing in project management, with AI-applications and tools such as predictive analytics, natural language processing, and machine learning becoming more widespread. Jannat et al. (2024) findings reveal also that AI positively influences decision-making,

reduces costs, and enhances overall project efficiency. Despite these advantages, challenges such as organizational resistance, limited expertise, and concerns about data privacy continue to hinder the broader adoption of AI in project management. According to Jannat et al. (2024), the adoption rate of artificial intelligence in project management was 97% in 2024, compared to only 10% in 2010.

However, integrating AI trustworthiness assessments into project management continues to be challenging, particularly for small and medium-sized companies (SMEs). Conducting trustworthy AI evaluations throughout a project requires practical-minded developers who embrace pluralistic values and understand the need to align the “*what*” with the “*how*”. In other words, how to implement a fully ethical approach to AI design, development, deployment, and use, as Morley et al. (2020) point out. For early adopters, this can create a competitive disadvantage, as it often necessitates additional training for developers or outsourcing specialized expertise, both of which introduce extra costs for the business. However, this is what the EU regulation requires from organizations developing high-risk artificial intelligence tools (EU AI Act, 2026).

As addressed, the increasing adoption of artificial intelligence in healthcare product development has introduced new layers of complexity into project management practices. While AI offers significant potential to enhance decision-making, accelerate development cycles, and improve resource allocation, it simultaneously introduces socio-technical risks that traditional project management frameworks are not designed to address, as Topol (2019) reports. As a result, integrating structured AI trustworthiness assessment methods, such as the Z-Inspection® process, into project management has become essential for ensuring that AI-enabled healthcare projects remain ethically sound, regulatory compliant, and operationally robust. This is also a requirement of the EU's Artificial Intelligence Act (EU AI Act, 2026).

Generally, any audit process by its nature, is slow, detailed, and methodical. These qualities often clash with the fast-paced development and innovation work that is typical for AI-driven technologies (Raji et al., 2020). Still, slowing down deliberately is essential as algorithms are introduced into increasingly high-stakes settings. Looking at historical cases across various industries, it becomes clear that audits can help anticipate negative

outcomes before they arise, like any other process in the project management pipeline. Auditing and other administrative processes also support decision-making around mitigation strategies, clarify and track potential adverse effects, and help identify harmful feedback loops and broader system-level risks. As Radi et al. (2020) emphasize, when carried out by a dedicated internal team, these audits function within the product development cycle and can even guide organizations toward discontinuing an AI project if its risks ultimately outweigh its benefits.

As described, trustworthy AI frameworks emphasize principles such as transparency, accountability, fairness, and human oversight. These principles align closely with the core responsibilities of project management, particularly in regulated domains such as healthcare. PMs are tasked with safeguarding quality, mitigating risks, and ensuring that project outcomes align with stakeholder expectations and regulatory requirements (Benjamins et al., 2020).

AI systems, however, introduce novel forms of risk: algorithmic bias, data governance challenges, model drift, and opaque decision-making. These require specialized assessment mechanisms. The integration of trustworthiness assessment into project management therefore represents a natural extension of existing governance structures rather than an external or optional addition (Benjamins et al., 2020).

As we describe, Z-Inspection[®] (Zicari et al., 2022) can provide a solution to this challenge, and it is one of the practical frameworks for assessing AI trustworthiness, and project management is one of the few fields that desperately requires structured trustworthy AI checkpoints. Integrating these two can provide an effective way to modernize project management in healthcare product development, and this integration also fulfills the notion of the EU regulation.

In this thesis, the Z-Inspection[®] method (Zicari et al., 2022) provides a structured, context-sensitive approach to evaluating the ethical and societal impact of AI systems. Its emphasis on socio-technical inquiry, stakeholder engagement, and iterative evaluation makes it particularly suitable for incorporation into the project lifecycle. When mapped onto established project management phases, Z-Inspection[®] can serve as a complemen-

tary governance mechanism that enhances both the rigor and the accountability of AI-related project activities and actions.

3.3 Artificial Intelligence Trends in Project Management

Project management continues to evolve as it responds to shifts in the business landscape and the introduction of new technologies that shape how projects are planned, carried out, and finalized. Emerging trends, such as the development trends of AI in project management, reflect these ongoing changes. As a key organizational and strategic function, project management supports goal achievement and enhances competitiveness in a global economy (Duică et al., 2024).

With advancements such as artificial intelligence, automation, and faster data-processing capabilities, project management methods are rapidly transforming. Likewise, progress in communication and collaboration, especially the rise of remote work and virtual teams has significantly altered how projects are organized and coordinated. For example, Duică et al. (2024) found out in their study that:

1. In 2020, it was reported that 98% of participants wanted the option to work remotely at some point in their careers, emphasizing the importance of adopting tools and strategies for effective remote project management (Duică et al., 2024).
2. It is forecasted that by 2030, the use of artificial intelligence will reduce the workload of project managers by 80%. This shift will result from AI taking over routine tasks such as data collection, monitoring, and reporting (Duică et al., 2024).

Several recent studies have examined how AI is being applied in project management (Müller et al. 2024). Current literature largely centers on three key themes: managing information, supporting decision-making, and the use of machine learning (Gil et al., 2021; Alevizos et al., 2023; Taboada et al., 2023; Salimimoghadam et al, 2025; Ibadildin et al., 2025; and Vergara et al., 2025). Common artificial intelligence applications in project management are illustrated in **Figure 5**.

The study by Ibadildin et al. (2025), for instance, examines by conducting a bibliometric review of academic literature and patents published between 2001 and 2024, how artificial intelligence is being incorporated into project management. The study's goal is to

highlight major trends, emerging uses, and worldwide contributions in this developing area. The Ibadildin et al. (2025) analysis draws on 51,752 patents from the *Lens* (<https://about.lens.org/>), 5.5 million scholarly entries from *Google Scholar* (<https://scholar.google.com/>), and 436,380 records from the *Web of Science* (<http://apps.webofknowledge.com/>), offering a broad picture of AI's influence on project management.

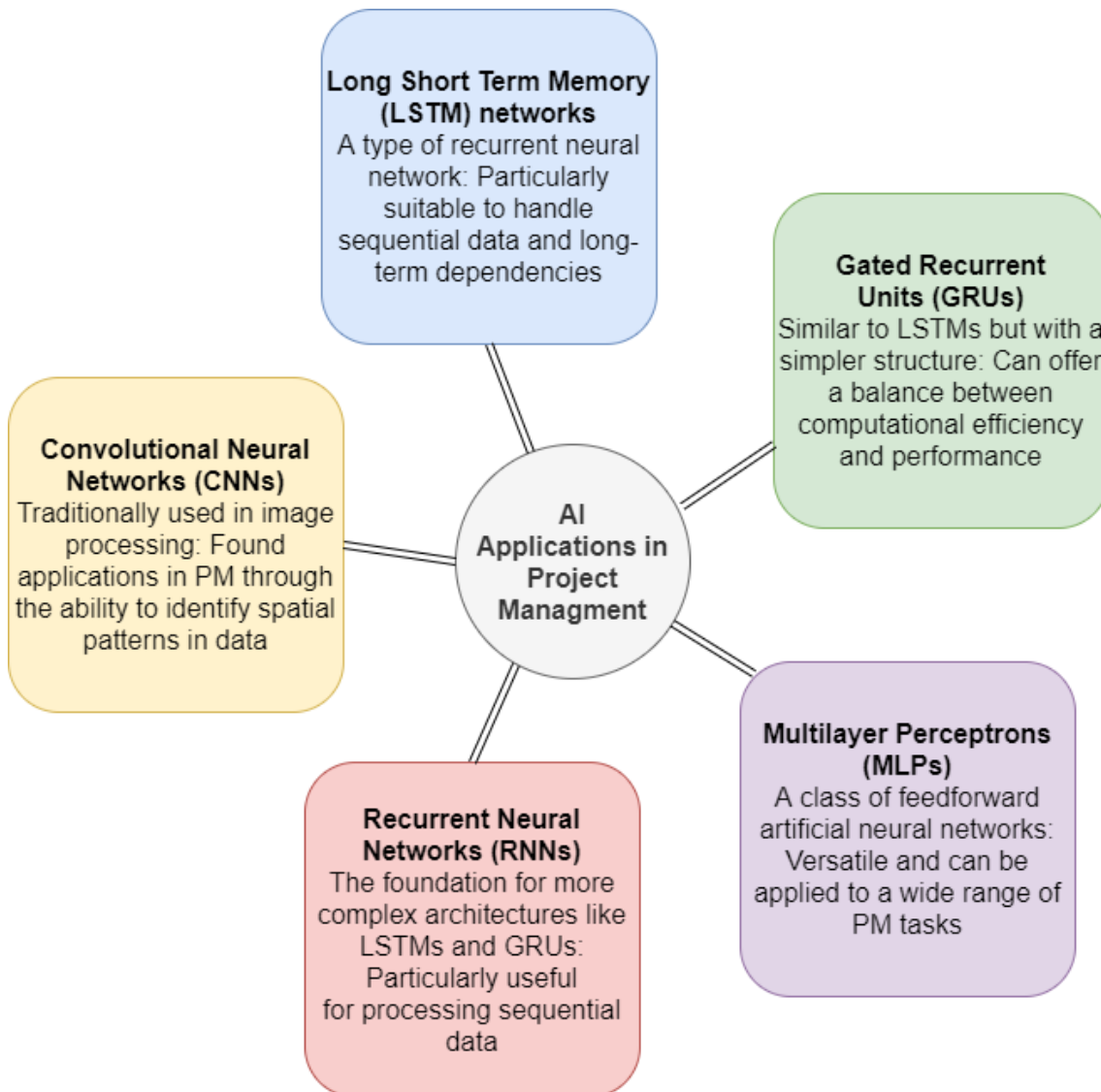


Figure 5: Common underlying technologies of artificial intelligence in project management. Adapted from Vergara et al. (2025).

Furthermore, Ibadildin et al. (2025) trace how AI-enabled project management practices have evolved, especially in areas such as resource allocation, risk assessment, and schedule optimization. The results show a significant rise in research related to AI and project management, with China and the United States producing the most output. A sharp increase in publications after 2019 suggests that both the COVID-19 pandemic and the growing push for digital transformation have accelerated interest in the topic.

Even with AI becoming more widely used in project management, the study (Ibadildin et al., 2025) identifies several *research gaps*, particularly regarding interdisciplinary approaches, real-world AI implementation, and ethical issues linked to algorithm-based decision-making. In a bibliometric analysis of innovation projects, Mesa Fernández et al. (2022) discover similar findings as Ibadildin et al. (2025). Furthermore, they note that AI methods are being applied in project management within dynamic environments to boost productivity. Their study (Mesa Fernández et al., 2022) gathered and examined key areas of technological innovation in project management, such as risk management, cost control, and schedule adherence, where AI techniques have the most significant influence. In other words, the studies reviewed mainly deal with traditional project management tasks that are not related to *assessing or regulating the trustworthiness* of artificial intelligence in project management process.

The study by Paporic & Bodea (2024) adds to existing research by offering insights into how generative AI (GenAI) is transforming the project management landscape and underscoring the need for new skills and roles to support this technological shift. It emphasizes how the responsibilities of project managers and other key stakeholders are changing in response to GenAI. Traditional positions, such as Project Manager, Team Member, and Business Analyst, are being reshaped by GenAI's capabilities, creating a demand for expanded skills and new competencies. Additionally, Paporic & Bodea (2024) emphasize the role of AI ethics, which becomes essential to ensure that AI systems adhere to ethical guidelines and address potential biases in project management tasks. We believe this is also something that needs to be taken into account in a trustworthy assessment of AI in product development in project management.

Felicetti et al. (2024) found in their study that when project managers use AI chatbots creatively, they not only apply the technology but also push its limits, resulting in *new and innovative ways* of managing projects. The results (Felicetti et al., 2024) provide project managers with guidance on how to adapt to and make effective use of generative AI tools, ensuring they are prepared to meet the changing demands of modern project management. The findings by Felicetti et al. (2024) offer organizations insights on how to harness Innovation attitude, peer influence, and task-related fit to support the *successful adoption and use* of GenAI in project management.

Hughes et al. (2025) remind us that the future of project management is being shaped not only by advances in AI but also by a growing emphasis on environmental sustainability. To remain relevant, project management practices must evolve to address the substantial ecological impacts that have often been ignored. Incorporating AI into project work offers a promising route to improve efficiency, agility, innovation, and resilience while simultaneously reducing environmental harm. Leveraging technologies such as AI and automation can lower energy consumption and provide greener alternatives to traditional project systems, according to Hughes et al. (2025).

In this thesis, we address the research gap related to the ethics of AI (Ibadildin et al., 2025). Furthermore, we incorporate AI trustworthiness assessment into project management processes in line with EU regulations – a research gap like that indicated, for example, by Mesa Fernández et al. (2022). We also use GenAI to gain new insights into our development work - the same GenAI skills are also required for modern project managers, as Paporic & Bodea (2024) suggest.

3.4 Impact of Artificial Intelligence on Project Management

Fridgeirsson et al. (2021) highlight that the contribution of project-based work to the Gross Value Added (GVA) - the core component of Gross Domestic Product (GDP) - of Germany, Norway, and Iceland increased by more than 20% between 2009 and 2019. This rise underscores the growing significance of project management and supports the argument that an organization's long-term sustainability is closely tied to its ability to manage change effectively.

The study by Hossain et al. (2024) underscores AI's strong potential to reshape project management by boosting efficiency, enhancing decision-making, and optimizing resource use. While certain challenges remain, they can be addressed through careful planning and dedicated training and development efforts. According to Hossain et al. (2024), the implementation of AI tools has resulted in significant reductions in project schedules by up to 20% and project costs by 15%, demonstrating their ability to improve overall project efficiency. Furthermore, the application of predictive analytics improved early risk identification by 30%, enabling a more proactive and informed approach to decision-making. Shoushtari et al. (2024) have observed similar impact trends in project management effectiveness when applying artificial intelligence tools (**Figure 6**).

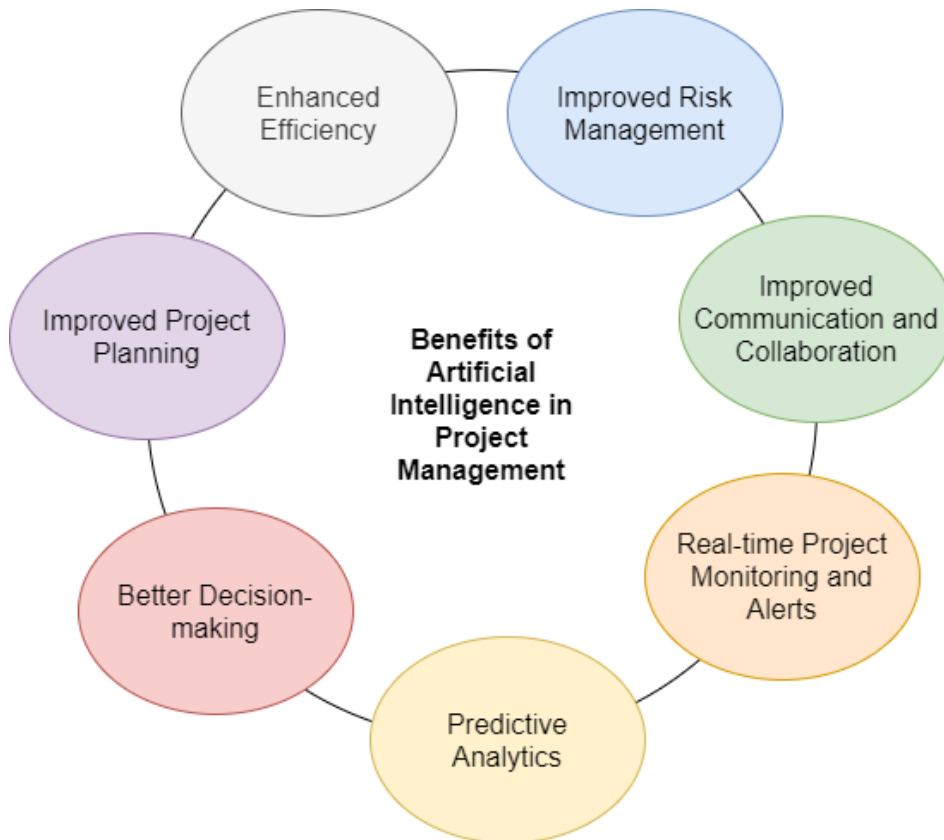


Figure 6: Benefits of artificial intelligence in project management, according to Shoushtari et al. (2024).

Similarly, the study of Jannat et al. (2024) reports findings from several case studies illustrating how AI has been applied in project management. Their study measures im-

provements in key performance metrics, including cost savings, time efficiency, quality, and safety, and highlighting the concrete advantages of integrating AI. In the healthcare sector specifically, the study reveals that using machine learning led to an 18% increase in resource optimization and a 30% rise in patient satisfaction within project management practices (Jannat et al., 2024).

The use of natural language processing (NLP) and intelligent resource allocation made the management of a healthcare project considerably more efficient, as Jannat et al. (2024) report as well. The NLP tools greatly enhanced communication within the team by enabling easier, real-time sharing and understanding of information. The project studied also improved the handling of patient data, which strengthened decision-making processes. Overall, Jannat et al. (2024) reported a 30% boost in efficiency, highlighting the substantial influence of artificial intelligence on managing complex healthcare projects. In summary, the benefits and impacts of using AI in project management, as shown in **Figure 6**, are obvious. However, in this thesis, we do not treat AI as a tool that helps project managers automate their work but rather embed trustworthy AI assessment into project management processes in AI-driven innovation and development work. We focus more on the ethics of AI, guided by the EU AI regulation, than on cost-effectiveness or administrative efficiency.

3.5 Risks, Opportunities, and Trustworthiness Concerns

Fridgeirsson et al. (2021) examine how AI could enhance the project management profession across all ten knowledge areas defined in the Project Management Institute's PMBOK framework (PMI, Chapter 4, 2021). They conducted a survey consisting of a panel participant of project management specialists, offering their perspectives on how AI is likely to influence the field over the next decade. Their responses strongly suggest that AI will become an integral component of project management practice and will shape each knowledge area in the near future. Furthermore, Hughes et al. (2025) point out that Analysts at *Gartner Group* have estimated that "*80 percent of current project management tasks will be eliminated by 2030*" as AI takes over several traditional project management tasks, such as gathering and processing data, monitoring project progress,

and potentially reshaping many of the core processes that support project and program management.

Ibadildin et al. (2025) conclude in their thoughtful bibliometric analysis that, although AI offers significant opportunities to improve project management, its adoption must be approached with care, ensuring that human-driven values and strategic guidance stay central. By doing so, organizations can fully leverage AI's capabilities while upholding ethical principles and effectively responding to the emerging challenges of an increasingly digital world.

Prasetyo et al. (2024) further revealed that successful AI adoption relies not only on technical factors, such as infrastructure, system integration, and data readiness, but also on leadership commitment, strategic alignment, adequate funding, workforce skills, and supportive organizational culture. According to Shang et al. (2023), Singaporean project experts also call for visionary foresight combined with timely government support and strategic organizational AI adoption plans. This is seen as an essential part of companies' preparation for a technologically AI-enabled future in project management.

Project managers now have a responsibility to ensure that AI is used ethically, addressing concerns such as data privacy, algorithmic bias, and transparency (Hughes et al., 2025). They play a vital role in establishing guidelines for responsible AI use and ensuring that AI initiatives reflect ethical principles and organizational values. According to the Hughes et al. (2025) study, this responsibility also extends to governance, risk management, and maintaining strong cybersecurity practices. Furthermore, Integrating AI requires strengthened leadership capabilities, continuous professional development, and strategic investment in training so that teams are equipped to manage complex systems that are sensitive to bias. Furthermore, project managers must increasingly act as ethical guardians of AI, applying governance standards, validating system outputs, and safeguarding stakeholder confidence through responsible oversight and thoughtful design. This thesis examines the above-described risks, opportunities, ethical principles and regulatory considerations in AI-based product innovation and development for project management practices. In addition, we present new perspectives on the field and highlight future research opportunities.

4 Machine Learning in Healthcare Innovation

4.1 Federated Machine Learning

In today's data-driven environment, healthcare is undergoing a major transition from traditional, centralized data systems toward distributed architectures that enable new technological possibilities. One promising approach within this shift is federated learning (FL), first introduced by Google in 2016 (Konečný et al., 2016). Federated machine learning (FML) uses machine learning (ML) and deep learning (DL) techniques in which models are trained locally on decentralized datasets, and only the learned model parameters, not the raw data, are shared and combined. The FML approach is especially valuable in the medical field, where safeguarding patient privacy, protecting sensitive information, and meeting strict regulatory standards are essential. FML offers an efficient way to utilize electronic health records (EHRs) from, e.g., multiple hospitals as a collaborative data platform while preserving confidentiality and avoiding the exchange of personal data (Rahman et al., 2023). This section gives a technical perspective to the needs of trustworthy AI assessment in healthcare project management in terms of FML.

Using federated machine learning, or FML, in healthcare can enhance the use of medical data, improve diagnostic precision, and maintain system safety. However, because federated machine learning can be considered a "*black box*," stakeholders may be concerned about whether its processes can be fully monitored throughout the lifecycle and whether the results can be made understandable and trustworthy (Badhon et al., 2025).

4.2 On Trustworthy Explainable Artificial Intelligence

Explainable artificial intelligence (XAI), also known as interpretable machine learning (IML), aims to enhance the transparency of AI as well as machine learning systems by using either inherently interpretable models or post-hoc explanation methods (Ducange et al., 2024). Several key factors encourage the integration of XAI into FML research and development efforts, as Albahri et al. (2023) describe:

Observability: Requesting explanations is a valuable educational approach for deepening understanding and clarifying the task at hand. Explainability enables researchers and practitioners to uncover new insights and verify the correctness of an AI system, ultimately resulting in more trustworthy outcomes.

Fairness considerations: Extensive research has documented biased or uneven results generated by certain AI systems, underscoring the importance of explanations to ensure algorithmic decisions are fair and free from discrimination.

Error reduction: Explanations serve not only to justify an AI system's outputs but also to help identify and prevent potential mistakes.

Performance enhancement: The ongoing pursuit of improving AI efficiency further supports the need for explainability. XAI methods are especially useful because they reveal the reasoning behind model outputs, helping users understand where adjustments can be made to boost performance.

The number of research publications on trustworthy explainable artificial intelligence (TXAI) in the healthcare sector has been clearly increasing. Between 2012 and 2015, only a small number of studies appeared. Publication activity began to pick up in 2017 and 2018, with one article released each year. This growth continued throughout 2019 and 2020, culminating in a notable rise to 13 publications in 2020. The trend persisted in 2021 and 2022, during which 24 and 21 articles were published, respectively, according to Albahri et al. (2023).

At the same time, as Albahri et al. (2023) report, insufficient trustworthiness in AI systems remains a major obstacle to their adoption in healthcare. When AI tools lack credibility, both patients and clinicians face greater uncertainty in their decision-making, which in turn complicates accountability for potential mistakes. As a result, the field is still striving to establish a clear gold standard for what constitutes trustworthy AI. According to the literature, achieving trustworthiness depends on three core principles: legitimacy, morality, and reliability or robustness.

Albahri et al. (2023) identify four primary objectives for implementing trustworthy AI in healthcare services:

- Enhancing patient experience.

- Supporting and improving overall population health.
- Lowering healthcare-related costs.
- Improving the working conditions and professional well-being of healthcare staff.

4.3 Foundations of Trustworthy Federated Machine Learning

To address the concerns above, the concept of trustworthy federated machine learning (T-FML, see **Figure 7**) has emerged, designed to align with the expectations of users, regulators, and other involved parties. In healthcare, T-FML strives not only to enhance patient care through collaborative AI methods but also to rigorously tackle essential issues such as security, privacy, fairness, and transparency (Li et al., 2023).

Zhang et al. (2024) divide trustworthy federated machine learning or T-FML into five main categories based on its core requirements. These include ensuring that user data is handled lawfully and remains private (privacy), that system deployment is secure, confidential, and free from errors (security), that model training remains dependable and consistent (robustness), that inputs from users or customers are treated impartially (fairness), and that the system's decisions can be clearly understood (explainability).

In general, Alzubaidi et al. (2023) report that the three most important requirements for trustworthy AI in healthcare are accuracy, explainability, and fairness. They also highlight that ethical considerations, scientific evidence, and issues related to bias and equity are essential for fostering trust between medical professionals and patients - while noting that privacy, security, and other technical concerns are not a focus in this context.

Furthermore, Alzubaidi et al. (2023) stress that, alongside ethical principles, strong evidence is vital in healthcare, particularly when confirming the validity of AI algorithms. Yet research in this area often uses inconsistent terminology and methods. Applying rigorous scientific standards during design and development, together with solid clinical proof that AI improves both effectiveness and safety, enhances the perceived trustworthiness of AI systems.

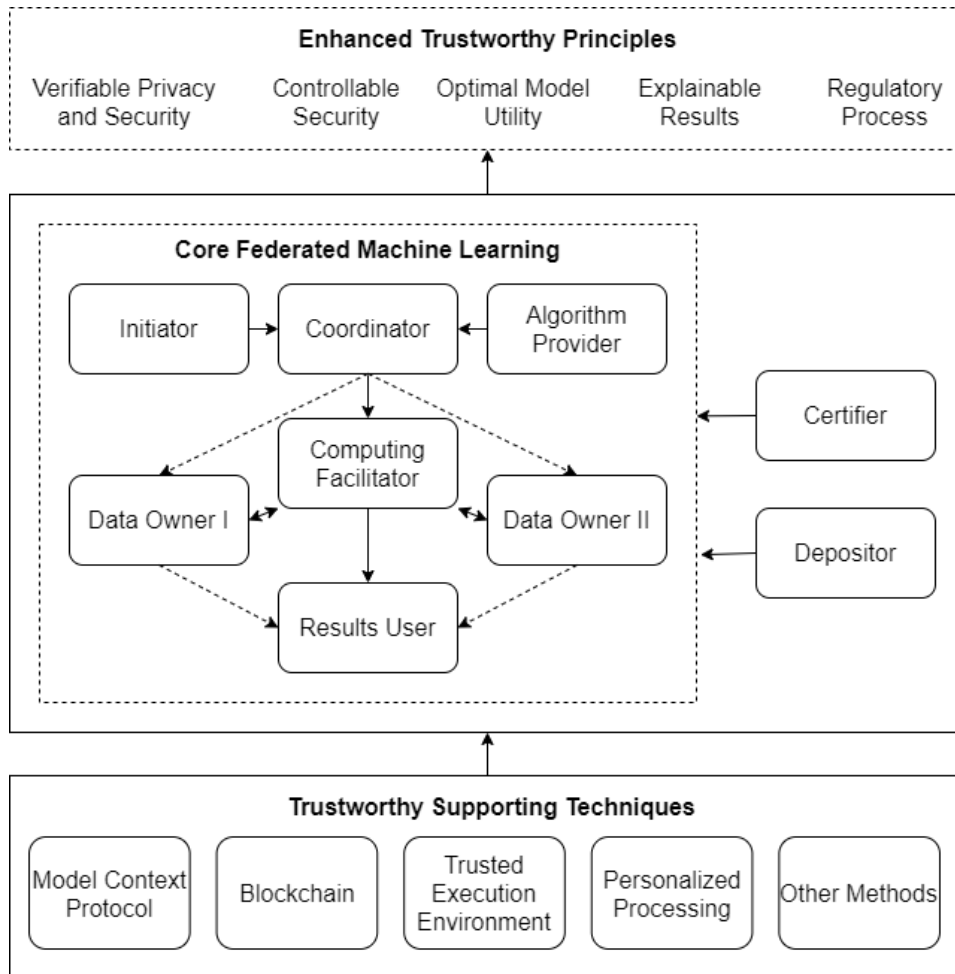


Figure 7: Framework for T-FML. Adapted from AISC (2024). Arrows with solid lines represent technology, data or information flows, while arrows with dashed lines represent control or management.

4.3.1 Privacy and Security

Data privacy and security concern the proper protection of sensitive information within an FML system to safeguard against unauthorized access or data leakage. Protecting both the data and the model parameters is crucial, as they may reveal confidential details about their owners. Any possibility of identifying an individual or household, whether directly or indirectly, must be clearly ruled out. Personal information such as names, ages, gender, facial photos, and biometric details like fingerprints all fall under this category. Because this information is valuable and subject to legal and regulatory safeguards,

strong privacy protection is a key aspect of establishing trust in an FML system (Zhang et al., 2024).

Healthcare data is naturally diverse, encompassing formats such as medical images, clinical documentation, and sensor-generated data, each bringing its own structural and interpretive challenges. Federated machine learning (FML) systems must be capable of handling this variety effectively while still maintaining strong model performance.

The preparation and refinement of data used in AI, such as cleaning, assessing its value, and annotating it, often require significant manual effort and have a substantial impact on the reliability of the resulting models. As Liang et al. (2022) emphasize, every phase of the AI data lifecycle from initial design to later adjustments and evaluation needs careful attention to strengthen the dependability of FML.

4.3.2 Accountability, Bias, Fairness, and Transparency

Fairness encompasses core ethical values, moral duties, and responsible practices. In the context of FML, it means ensuring that systems do not exhibit prejudice, bias, or discriminatory behavior toward any group or individual. According to Zhang et al. (2024), fairness can be approached through three dimensions:

1. **Performance fairness**, which aims for consistent accuracy levels across all participating clients.
2. **Input fairness**, which allocates rewards or influence based on each client's actual contribution; and
3. **Attribute fairness**, which ensures that individuals with similar characteristics receive similar predictions and that sensitive attributes, such as gender or age, do not lead to biased outcomes.

Bias in algorithms can significantly threaten fairness in the distribution of healthcare resources. When implemented effectively, FML can help reduce such biases by incorporating diverse datasets from different institutions and population groups.

Transparency and accountability present additional challenges for FML. Because the approach avoids the direct exchange of training data, it becomes difficult to inspect or verify the training process. This makes it harder to guarantee that the model remains unbi-

ased and has not been manipulated. As a result, developers may struggle to meet regulatory requirements that demand transparency or auditability as conditions for deploying AI systems, as Tariq et al. (2024) note.

Although related to interpretability and explainability, transparency specifically concerns how clearly the functioning and decision-making processes of an AI system can be understood by stakeholders. This clarity is especially important in healthcare, where trust is essential (Tariq et al., 2024).

4.3.3 Explainability and Interpretability

The principle of transparency is vital to ensure that ML and DL models of FML are not just black boxes, but systems whose decision-making processes can be questioned and understood by clinicians and other stakeholders.

Explainability refers to how well users can interpret and articulate how an FML system arrives at its decisions and results. It includes two core components: ante hoc explainability and post hoc explainability. Ante hoc explainability focuses on ensuring transparency and clarity in the system's design from the outset, aiming to evaluate the interpretability of each step in the system's workflow. Post hoc explainability, on the other hand, involves using external methods or tools to shed light on the model's decisions after they have been generated. This helps users assess whether the system's output can be meaningfully understood. Strong explainability enhances user trust and improves the usability of the system across different types of users, supporting more reliable and effective human-AI interaction, as Zhang et al. (2024) clarify.

While explainability refers to making the model understandable afterwards, interpretability refers to how directly a human can understand the inner workings of an FML model. In the highly sensitive context of healthcare, interpretability can literally be a matter of life and death. For example, using techniques such as XAI, researchers are developing models that provide insight into the logic and predictions of the models, which foster trust not only among healthcare professionals but also among patients, who are the ultimate beneficiaries of improved treatment protocols, as Rahman et al. (2023) write.

For nuances of terminology, Markus et al. (2021) examine the definitions of various AI-related terms, such as explainability, interpretability, comprehensibility, intelligibility, transparency, and understandability, within the healthcare context. Their review suggests that additional measures, such as thorough reporting of data quality, extensive external validation, and regulatory oversight, may be required to ensure AI systems in healthcare are genuinely trustworthy.

4.3.4 Technical Robustness

Resistance or robustness to attacks are major challenges for FML. Robustness describes an FML system's ability to remain stable even under adverse or hostile conditions, which is essential because these systems typically operate in complex and unpredictable real-world settings. Robustness heavily influences how well an FML system performs in practice. Insufficient robustness can result in unintended or harmful outcomes and reducing trust in the system. For this reason, it is important that aggregated models remain accurate, fair, and capable of generalizing across diverse data sources (Zhang et al., 2024).

When deploying FML systems, it is essential to evaluate their robustness. Users need to know under which conditions the system's predictions can be trusted or dismissed. Which kinds of populations the system is suitable for, and what limitations exist in the data on which the system was trained (Cutillo et al., 2020). Thus, it is not uncommon for FML systems to exhibit vulnerabilities, including susceptibility to model-poisoning, data-poisoning, and inference attacks (Guo, J. et al., 2024).

Chen et al. (2025) outline both unintentional vulnerabilities and deliberate attacks, along with corresponding attack strategies. To mitigate these risks, they assess a range of existing defense techniques and stress the need for a comprehensive risk management framework.

4.3.5 Data Integrity, Regulatory Compliance, and Trust

Concerns related to data integrity, privacy, and security are receiving growing attention. These concerns make it essential for FML researchers and developers to follow all legal requirements when collecting, handling, and utilizing data.

A wide range of laws and regulations have been introduced to protect the security and privacy of raw data. Examples include China's Cybersecurity Law (CLPR), the European Union's General Data Protection Regulation (GDPR, 2016), the California Consumer Privacy Act (CCPA), and the Consumer Privacy Bill of Rights (CPBR). According to Chen et al. (2025), these privacy rules are estimated to apply to as much as 75% of the global population, with more than 80% of companies worldwide expected to be compliant by 2023. Trust is closely tied to regulatory compliance. When federated machine learning practices align with established legal standards, they enhance both the legitimacy and the perceived reliability of these technologies in healthcare. That is to say, the medical field must create clear guidelines for emerging AI technologies and re-evaluate the existing regulatory environment to ensure that clinicians and patients can have confidence in these systems, as the EU AI Act (2026) requires.

This guarantees that trustworthy FML in healthcare is grounded in evidence, transparent, free from bias, and supportive of equity. Because technological advancements progress rapidly, healthcare professionals must continually adapt. This requires well-defined responsibilities for AI developers and researchers, as well as for healthcare institutions and leaders who oversee the assessment of AI tools in clinical settings. These actors should establish trustworthy processes that uphold ethical standards, scientific evidence, and equity in practical applications. Meaningful collaboration among diverse stakeholders is vital for addressing these challenges early in the implementation process, as Albahri et al. (2023) note.

Although FML restricts the sharing of raw training data among participants, it also brings its own set of security risks. Clients must be able to trust the central aggregator to provide secure implementation, interact only with verified participants, follow the established aggregation protocol, and use the final model solely for agreed-upon purposes.

Likewise, the aggregator must trust clients to supply high-quality data, follow the training protocol, and respect the model's intellectual property rights (Guo, J. et al., 2024).

All these trustworthy conditions can be difficult to satisfy in real-world environments, particularly in competitive contexts where participants might weaken their own data protections or collaborate in ways that compromise the system.

4.4 Technologies Supporting Trustworthy Federated Machine Learning

From a technical point of view, Zhang et al. (2024) explain that FML can be organized according to two main communication structures. In centralized federated machine learning (CFML), a central server or aggregator coordinates the process by repeatedly exchanging global model updates and local model adjustments with participating clients. In contrast, distributed federated machine learning (DFML) enables clients to collaborate directly by sharing model parameters with one another in a peer-to-peer manner.

Communication within an FML system typically relies on two layers: the control plane and the data plane. Hypertext Transfer Protocol (HTTP) and Message Queuing Telemetry Transport (MQTT) serve as established standards for these interactions, with HTTP supporting request, response communication and MQTT enabling publish, subscribe mechanisms. MQTT's asynchronous, subscription-based model provides a notable advantage over the more synchronous structure of HTTP (Boiano et al., 2024).

FML can also be categorized based on how data is partitioned across participants:

1. **Horizontal federated machine learning (HFML):** clients share the same feature space but possess different data samples.
2. **Vertical federated machine learning (VFML):** clients have data on the same individuals but with different feature sets.
3. **Federated transfer machine learning (FTML):** both the samples and the features differ across clients.

The infrastructure layer of an FML system encompasses resource allocation, training data management, and the underlying communication framework.

According to Rahman et al. (2023), the key stages of an FML workflow include:

- System setup and client selection.

- Local model training and updating generation.
- Model aggregation and distribution, using methods such as Google's Federated Averaging (FedAvg).

4.4.1 Internet of Medical Things

Nguyen et al. (2022) review the use of FML in smart healthcare settings, particularly within the Internet of Medical Things (IoMT). The IoMT consists of devices, such as wearable sensors, that gather health-related data for AI-driven analysis. Their proposed framework incorporates factors such as resource efficiency, improved security and privacy, incentive mechanisms, and personalization. These are aligned with principles found in explainable AI (XAI). Overall, existing research indicates that federated machine learning offers promising approaches for enhancing privacy protection in smart healthcare environments, as Ali et al. (2022) state.

4.4.2 Blockchain Architecture

According to the review by Albahri et al. (2023), blockchain architecture offers significant benefits for tracking data provenance and monitoring the model training process, helping to detect performance deterioration or dishonest activity. By establishing a transparent and hack-proof environment, blockchain can strengthen trust among participants and support large-scale collaboration. This is an essential factor in building trustworthy AI systems relying on vast datasets. This also makes blockchain an important element in assessing how federated learning could serve as a future foundation for digital healthcare.

For example, Yang et al. (2022) developed a blockchain-based T-FML architecture to ensure security and privacy, which utilizes secure global aggregation and blockchain technology to resist attacks from malicious edge devices and servers. Similarly, Stephanie et al. (2022) proposed a secure, multi-party ensemble federated learning approach using blockchain technology, which enables collaborative learning of heterogeneous models from healthcare institution data without violating user privacy. Blockchain features also

enable data integrity across parties without a centralized server, while providing auditability and version control capabilities for each healthcare institution.

4.4.3 Trusted Execution Environment

A different route to achieving privacy in machine learning is the use of confidential computing. This approach allows code and data to run securely even in environments that cannot be fully trusted, such as public cloud platforms, by relying on hardware-based trusted execution environments (TEEs). In contrast to federated machine learning, this method places substantial trust in the TEE hardware to safeguard data stored and processed in a potentially risky remote setting, as Guo et al. (2024) describe. Still, TEEs face limitations when applied to FML, including restricted storage capacity and exposure to side-channel attacks, as Zhang et al., (2024) write.

Importantly, however, FML and confidential computing should not be viewed as opposing solutions. With thoughtful system design, they can be integrated to leverage the strengths of both. This combined model keeps sensitive information within its secure domain while maintaining transparency and accountability. So-called confidential federated learning (CFL) can therefore reduce the risk of large-scale attacks on FML, support broader adoption in sectors with strict privacy requirements, and help organizations meet emerging AI regulatory standards, as Guo et al. (2024) suggest.

5 Methods

In general, the main goal of this thesis is to produce actionable knowledge that supports professionals in designing solutions to context-specific problems. To achieve this, we use a design science research approach, which emphasizes the systematic development and evaluation of artifacts (Bayazit, 2004). This approach is particularly suitable for structuring and evaluating feasible and useful methods for integrating trustworthy AI assessment practices into project management processes. As part of our design science methodology, the thesis also includes a strong qualitative research perspective.

Our practical work is mainly based on four-month experience in the FEHLS project, which was implemented at Tampere University in the summer of 2025. In order to gain the necessary understanding of current trends and applications of artificial intelligence in project management practices, we conducted an extensive literature review on the reported topics. In addition to our research work and literature review, we used prompt engineering design to simulate possible interview responses from human experts to support our ideas and observations. In our opinion, this can provide a solid and sufficiently broad background for the proposed framework. We believe that the methodology used in this study also provides a comprehensive and detailed approach to our thesis within a limited time frame and resources.

5.1 Trustworthy AI Impact Assessment in the FEHLS Project

The following sections provide an overview of the research methods in the FEHLS project. More detailed explanations can be found in the article, written by the developers and Z-Inspection® assessors of the FEHLS project (Kringen et al., 2026).

5.1.1 Study Design

The Federated Ethical Healthcare Learning Sandbox or FEHLS project, funded by *Nordic Innovation*, was organized as a coordinated cross-national collaboration between Den-

mark, Finland, and Norway (FEHLS, 2026). From a project-management perspective, the work was structured into two parallel tracks:

- Technical development of two interoperable machine-learning pipelines.
- Independent trustworthy AI assessment using the Z-Inspection® methodology.

The technical teams in Denmark (Aalborg University/RUGP) and Norway (NTNU/St. Olav's Hospital) built predictive models estimating the 3-month risk of a treatment for headache disorders failure using a federated machine learning pipeline. Each predictive model used locally extracted patient features and produced a continuous risk score. The assessment teams (Arcada University of Applied Sciences and Tampere University) operated separately to ensure unbiased evaluation (Happonen et al., 2025).

A key project management decision was to keep patient data within national borders. Only model parameters were exchanged, enabling each site to develop, test, and refine its pipeline independently while still contributing to a shared FML model. This setup allowed the consortium to test both the practical feasibility of FML and the trustworthiness of the resulting prototype in a controlled, cross-national clinical environment, as reported in Kringen et al. (2026).

5.1.2 Data Sources and Processing

Danish data came from general-practice (GP) electronic health records (over 64,000 patients), from which a cohort of 106 eligible headache patients was derived. Norwegian data (NorHead) included neurology outpatient documentation, yielding 454 patients for modelling. Both countries used Python-based extraction pipelines to convert raw clinical data into structured variables. Diagnostic and medication codes were harmonized to create a shared feature set (e.g., age, sex, common diagnoses, and frequently prescribed medications). A specific treatment for headache disorders itself was excluded from predictors to avoid data leakage (Kringen et al., 2026).

5.1.3 Development of Federated Machine Learning Model

Given limited and heterogeneous datasets, the FEHLS project consortium selected gradient boosting as the primary model in FML, with logistic regression as a baseline. Both sites in Denmark and Norway followed a standardized workflow: splitting data into training, testing, and validation sets. The workflow included also tuning hyperparameters via randomized cross-validation and evaluating performance using average precision and confusion matrices. In the FML setup, only model updates were shared and each team tested both local and FML models on its own data and the partner's data to assess generalizability, as Kringen et al. (2026) describe.

5.1.4 Design of Trustworthy Artificial Intelligence Assessment

The trustworthy AI assessment followed the three-phases of the Z-Inspection® methodology running in parallel with technical development. An independent, multidisciplinary team validated legal and ethical prerequisites, co-designed socio-technical scenarios with developers, and analyzed these using the Claims-Evidence-Reasoning framework (McNeill et al., 2011). Issues and tensions were mapped to the EU's AI HLEG requirements for Trustworthy AI producing actionable recommendations and a structured barrier assessment (Kringen et al., 2026).

5.1.5 Trustworthy Assessment Extensions

To strengthen risk governance, identified issues were also mapped to the Massachusetts Institute of Technology (MIT) AI Risk Repository (2026). Each issue was translated into a concise risk statement, categorized by likely harm pathways, and validated through consensus among assessors. This mapping complemented the barrier assessment by highlighting missing safeguards and proposing targeted mitigation measures, as Kringen et al. (2026) show.

5.2 Design of the Proposed Project Management Framework

We have explored the current state of artificial intelligence literature in project management. To our knowledge, there are little or no previous academic studies on the integration of trustworthy AI assessment and project management processes into a unified framework for healthcare innovation.

In this thesis, our project management views are based on the author's experiences from the FEHLS project and research in the field, as well as knowledge based on the author's previous studies and work experience. All this know-how together has given the basis for the proposed framework.

To support our design work and findings, we used prompt engineering with *ChatGPT* and *Microsoft 365 Copilot* (Giray, 2023) for the design of the proposed project management framework. This allowed us to confidently explore potential new avenues in the fields of project management and trustworthy AI, as Ye et al. (2024) encourage in their prompt engineering study.

Since integrating trustworthy AI assessment, such as the Z-Inspection® method, into project management processes is a new research perspective and addresses a clear research gap in the academic literature, we believe that chatbot responses can provide as fruitful insights into the integration framework of these two fields as human experts could provide – or even better, in this particular emerging multidisciplinary topic. In other words, finding external experts in this innovative field was considered challenging due to the limited time and scope of the thesis standards.

The following themes and related questions were used to elicit simulated expert opinions using the prompt engineering design technique in developing the proposed framework. The simulated responses obtained were analyzed to reflect on and improve our own original ideas. The themes and interview questions are as follows:

- 1) **Background and Context:** These questions help us to understand the interviewee's role and experience.
 - I. Could you describe your role and responsibilities in healthcare product development?

- II. What kinds of AI systems or data-driven tools have you encountered in your work?
- III. How would you describe your familiarity with AI governance or trustworthiness concepts?

2) AI Use in Project Management: These questions explore how AI is currently integrated into project management processes.

- I. In your experience, how is AI currently used in project management activities within healthcare product development?
- II. Which project management tasks do you believe are most affected by AI tools (e.g., planning, risk management, resource allocation, documentation)?
- III. What benefits have you observed or expect from using AI in project management?
- IV. What challenges or limitations have you encountered?

3) Perceptions of Trustworthy AI: These questions probe the interviewee's understanding of trustworthiness.

- I. When you think about "*trustworthy AI*," what aspects or principles come to mind?
- II. How important do you consider issues such as transparency, explainability, fairness, or human oversight in AI systems used in healthcare projects?
- III. Have you encountered situations where trustworthiness concerns influenced decisions about AI adoption or use?

4) Risks and Ethical Considerations: These questions connect directly to the Z-Inspection® focus on socio-technical risks.

- I. What kinds of risks do you associate with using AI in healthcare product development projects?
- II. How do you think AI might influence decision-making quality or accountability in project teams?
- III. Are there ethical or societal concerns you believe should be considered when integrating AI into healthcare project workflows?

- IV. How do regulatory requirements (e.g., MDR, EU AI Act) shape your view of AI risks?

5) Integration of Trustworthiness Assessment into Project Management: These questions address the core research question.

- I. Do you think trustworthiness assessment should be part of project management processes? Why or why not?
- II. At which stages of the project lifecycle (initiation, planning, execution, monitoring, closing) do you think AI trustworthiness should be evaluated?
- III. What practical challenges might arise when integrating trustworthiness assessments into project workflows?
- IV. What kinds of tools, frameworks, or guidelines would help project managers evaluate AI trustworthiness effectively?

6) Z-Inspection® Method (if interviewee is familiar): If the interviewee knows the method; if not, you can briefly explain it before asking.

- I. Are you familiar with the Z-Inspection® method for assessing trustworthy AI?
- II. Based on your understanding, how applicable do you think Z-Inspection® is to healthcare product development projects?
- III. What strengths or limitations do you see in using Z-Inspection® within project management?
- IV. How might such a method support compliance, risk management, or decision-making?

7) Organizational Readiness and Capability: These questions explore maturity and practical feasibility.

- I. How prepared do you think your organization (or the industry in general) is to assess AI trustworthiness systematically?
- II. What competencies or roles do you think are needed to support trustworthy AI assessments in projects?
- III. What barriers (cultural, technical, regulatory, resource-related) might hinder adoption?

8) Future Outlook: These questions help you capture forward-looking insights.

- I. How do you see the role of AI evolving in project management for healthcare product development?
- II. What changes do you think will be necessary to ensure AI is used responsibly and effectively?
- III. If you could recommend one improvement to current practices regarding AI governance in projects, what would it be?

9) Closing

- I. Is there anything else you would like to add that we have not covered?
- II. Are there documents, processes, or examples you think would be useful for my research?

The above questions were answered by *Microsoft 365 Copilot* using prompt engineering techniques (Ye et al., 2024).

6 Results

As stated, this thesis consists mainly of design and qualitative science research, as well as an extensive academic literature review. The core qualitative part, i.e. the multidisciplinary interviews used with the Z-Inspection® method, is reported in detail in the article by Kringen et al. (2026). In addition to the design, qualitative and literature review, chatbot interview responses enrich our findings for the resulting integration framework.

6.1 Findings of Trustworthy AI Assessment in the FEHLS Project

Within the FEHLS project, the two AI systems being developed were combined into an FML infrastructure designed to produce personalized treatment recommendations for patients experiencing migraines or headaches. Through the Z-Inspection® assessment, the project team showed that the FML architecture *successfully maintains data sovereignty* while *ensuring consistent ethical standards* across national boundaries, as reported in Kringen et al. (2026).

Early findings from the FEHLS project also indicate that Z-Inspection® can help uncover inflated or *unfounded developer claims*, *reveal challenges in collaboration across organizations*, and *identify evidence and mitigation strategies* to enhance the trustworthiness of healthcare-related FML systems. These insights shed light on how *EU regulations concerning ethical AI and medical product development* can be embedded into national regulatory processes, and how Z-Inspection® could serve as a core component of *AI product management practices* (Kringen et al., 2026).

In **Appendix 1**, claims, evidence, and reasonings identified by the multidisciplinary team during the FEHLS project's Z-Inspection® process are presented. The next sub-sections provide a short summary of the findings. More detailed preliminary results and their analysis can be found in the report by Kringen et al. (2026).

6.1.1 Priority Actions

Using the Claims-Evidence-Reasoning (CER) framework (McNeill et al., 2011), the FEHLS project identified practical trustworthiness issues in the federated machine learning setup. Although FML protects data privacy by design, several technical and organizational risks remain. By translating CER findings into a barrier assessment, the project highlighted where safeguards must be strengthened before the system can be considered trustworthy.

Four priority actions emerged in the study (Kringen et al., 2026):

1. Clarify the clinical meaning of similarity-based FML predictions.
2. Provide (technical) explanations to clinicians so they can understand and utilize them.
3. Define traceability and audit mechanisms that protect privacy while enabling medically justified identification.
4. Establish a lightweight but formal ethics and governance onboarding process that continues into deployment.

Table 2 in **Appendix 1** summarizes each issue, its technical implications, and recommended organizational actions, as discovered in the FEHLS project.

6.1.2 Identified Core Assessment Challenges and Recommendations

The Danish and Norwegian teams operated at different technology readiness levels or TRLs (Sauser et al., 2006) affecting their ability to benchmark and aggregate models. This mismatch complicated FML integration because each site had different pipeline maturity and evaluation practices. The governance team addressed this by aligning expectations and clarifying project aims (Kringen et al., 2026).

The assessment found demographic imbalance, e.g., overrepresentation of women, missing data, and cohort-specific characteristics that could introduce algorithmic bias. These issues reduce statistical reliability and may lead to poor predictions for underrepresented groups. Recommended actions include subgroup fairness testing,

weighting or calibration, and transparent model cards documenting limitations, as Kringen et al. (2026) describe.

Although SHapley Additive exPlanations or SHAP (Li, 2022) values and probability scores were available, clinicians may struggle to interpret them. This is a usability, not accuracy, problem. The project recommends a tiered XAI workflow: simple summaries for clinicians and detailed explanations for technical experts, supported by user testing and training (Kringen et al., 2026).

Logging and documentation practices were not fully specified in our study. In FML systems, stronger audit trails improve accountability but increase privacy risks. The assessment recommends privacy-preserving traceability, such as pseudonymized IDs, minimal audit fields, strict access control, and short retention periods (Kringen et al., 2026).

Differences in electronic health records (EHRs) formats, coding systems, and Application Programming Interfaces or APIs created friction in the federated machine learning workflow and increased preprocessing demands. To improve scalability, the project recommends a shared data dictionary, a minimal common feature set, and a lightweight integration layer across sites (Kringen et al., 2026).

FML models inherit errors from clinical records. Continuous monitoring, clinician review of outliers, and transparency about what the model can and cannot correct are needed to prevent “*garbage-in, garbage-out*” effects, as Kringen et al. (2026) point out.

The FML system lacked stress testing for rare events, such as adversarial inputs and robustness. Undetected malicious or erroneous data at local sites could compromise the aggregated FML model. A phased validation plan including synthetic edge cases, latency tests, and local pilots is recommended to strengthen reliability and support TRL progression, as found in Kringen et al. (2026).

Mapping the findings to the MIT AI Risk Repository placed the FEHLS project’s challenges within a broader risk taxonomy. Four domains were found especially relevant:

AI system safety and limitations: Low TRLs and limited validation increase the risk of unexpected behavior.

Human-computer interaction and discrimination: Model choice and explainability affect fairness and usability.

Transparency, privacy, and security: Documentation gaps and unclear contestability mechanisms mirror known risks.

Governance and organizational readiness: Early-stage deployment may influence clinical workflows and trust.

Although adversarial misuse was not a major concern at this stage, the repository highlights the need to anticipate such risks before scaling. Overall, the mapping helped convert abstract risks into concrete preventive and mitigative barriers, as reported by Kringen et al. (2026).

6.1.3 Strengths of Z-Inspection®

A key strength is the demonstration that a FML system can operate across the Nordic countries under the EU's GDPR and national data-sovereignty constraints. The FEHLS project shows that sensitive clinical collaboration is feasible without moving patient data across national borders and jurisdictions (Kringen et al., 2026).

Our trustworthy AI assessment confirmed that federated machine learning does not automatically guarantee fairness, transparency, or accountability. The Z-Inspection® approach, however, can help to surface these gaps early and translate them into practical governance measures, such as privacy-preserving audit logs and ongoing monitoring. Combined with regulatory frameworks like the EU AI Act, this dual-track approach can support a shift from minimal compliance toward genuinely trustworthy AI, as Kringen et al. (2026) argue.

Based on these findings, we can conclude that this part of the study answers our **Research Question 1 (RQ1)** and supports its hypothesis (**RH1**):

“The Z-Inspection® method provides measurable improvements in the robustness, transparency, and ethical compliance of trustworthy AI assessments compared to traditional assessment methods.”

6.2 The Proposed Project Management Framework

At this point, we would like to emphasize that modern project managers (PMs) need digital intelligence to be able to manage and execute projects efficiently and effectively (Marnewick & Marnewick, 2021). Digital intelligence skills of PMs play a significant role, especially in international projects, to ensure smooth and predictable project management.

In this thesis, we examine five general steps in the project management workflow for the proposed framework, as presented in the next sections.

6.2.1 Initiation Phase

During the initiation phase, Z-Inspection® supports early-stage feasibility analysis by clarifying the intended use of the AI system, identifying potential ethical and societal risks, and assessing alignment with healthcare regulations. This early integration ensures that trustworthiness considerations inform the project's foundational decisions, including scope definition and stakeholder analysis.

Project managers should pay particular attention to the following:

- Define the intended use, users, and context of the AI system.
- Identify high-level risks (e.g., fairness, privacy, safety, robustness).
- Document assumptions and constraints.

At this phase of the project, PMs set the ethical path for the project. As always, careful planning is very important.

6.2.2 Planning Phase

In the planning phase, trustworthiness criteria can be embedded into project plans, risk registers, and quality management strategies. This includes defining requirements for explainability, data governance, and human oversight. The very elements that are increasingly necessary for compliance with emerging regulatory frameworks such as the EU AI Act.

Project managers should pay particular attention to the following:

- Integrate trustworthy AI requirements into the project plan.
- Assign roles: who owns fairness testing, who handles model monitoring, who signs off on risk mitigations, etc.
- Choose frameworks (e.g., NIST AI RMF, ISO/IEC 42001, internal governance models, etc.). We, of course, recommend the Z-Inspection® methodology.

This is the phase where trust becomes operational.

6.2.3 Execution Phase

As the project moves into execution, Z-Inspection® provides mechanisms for continuous monitoring of ethical and technical risks. This is particularly important in healthcare contexts, where AI systems must be validated not only for technical performance but also for their impact on clinical workflows, patient safety, and organizational processes (He et al., 2019).

The Z-Inspection® method's emphasis on contextual evaluation ensures that AI behavior is assessed within the real-world environments in which it will operate.

For project managers, worth noting here are the following key issues:

- Conduct model evaluations: bias testing, robustness checks, data quality assessments.
- Implement human-in-the-loop controls where needed.
- Track risks in the same way you would track project economy or security issues.

In the execution phase, trust becomes measurable.

6.2.4 Monitoring and Controlling Phase

In the monitoring and controlling phase, Z-Inspection® contributes to ongoing quality assurance by enabling periodic trustworthiness audits, reassessment of risks, and documentation of compliance-related evidence. These activities strengthen the project's ability to respond to evolving risks, including model drift or changes in regulatory expectations.

Here, the key things to note by PMs are particularly:

- Continuous monitoring for drift, misuse, or unexpected outcomes.
- Periodic reassessment of risks as the system evolves.
- Transparent documentation of decisions and changes.

This is the phase where trust becomes sustainable.

6.2.5 Closing Phase

Finally, in the closing phase, Z-Inspection® supports the formal evaluation of the AI system's trustworthiness and the development of post-deployment monitoring plans. This ensures that ethical and societal considerations remain part of the system's lifecycle beyond the project's formal completion. For healthcare organizations, this is particularly valuable, as AI systems often require continuous oversight to maintain safety, reliability, and regulatory compliance.

Things project managers should take care of are in particular:

- Final trustworthy AI assessment.
- Lessons learned for future projects.
- Documentation for audits, regulators, or internal governance.

In the closing phase, trust becomes institutional knowledge.

6.2.6 The Big Picture of the Proposed Framework

The integration of the Z-Inspection® method into project-management processes can be understood as a structured, iterative governance workflow that embeds ethical and socio-technical evaluation into the lifecycle of AI system development. Z-Inspection®, as a methodology for assessing trustworthy AI, emphasizes contextual inquiry, multidisciplinary reflection, and evidence-based evaluation. When aligned with project-management practices, it becomes a systematic mechanism for ensuring that ethical, legal, and societal considerations are treated as core project deliverables rather than external add-ons.

A defining feature of Z-Inspection® is its reliance on multidisciplinary expert groups. When embedded into project management, this becomes a formalized governance structure:

- Ethical experts, domain specialists, engineers, legal advisors, and end-users participate in structured assessment sessions.
- Their insights feed directly into project decision-making.
- The project manager becomes the orchestrator of this cross-functional dialogue.

This transforms ethical reflection from an ad-hoc activity into a repeatable project process, as shown in **Figure 8**.

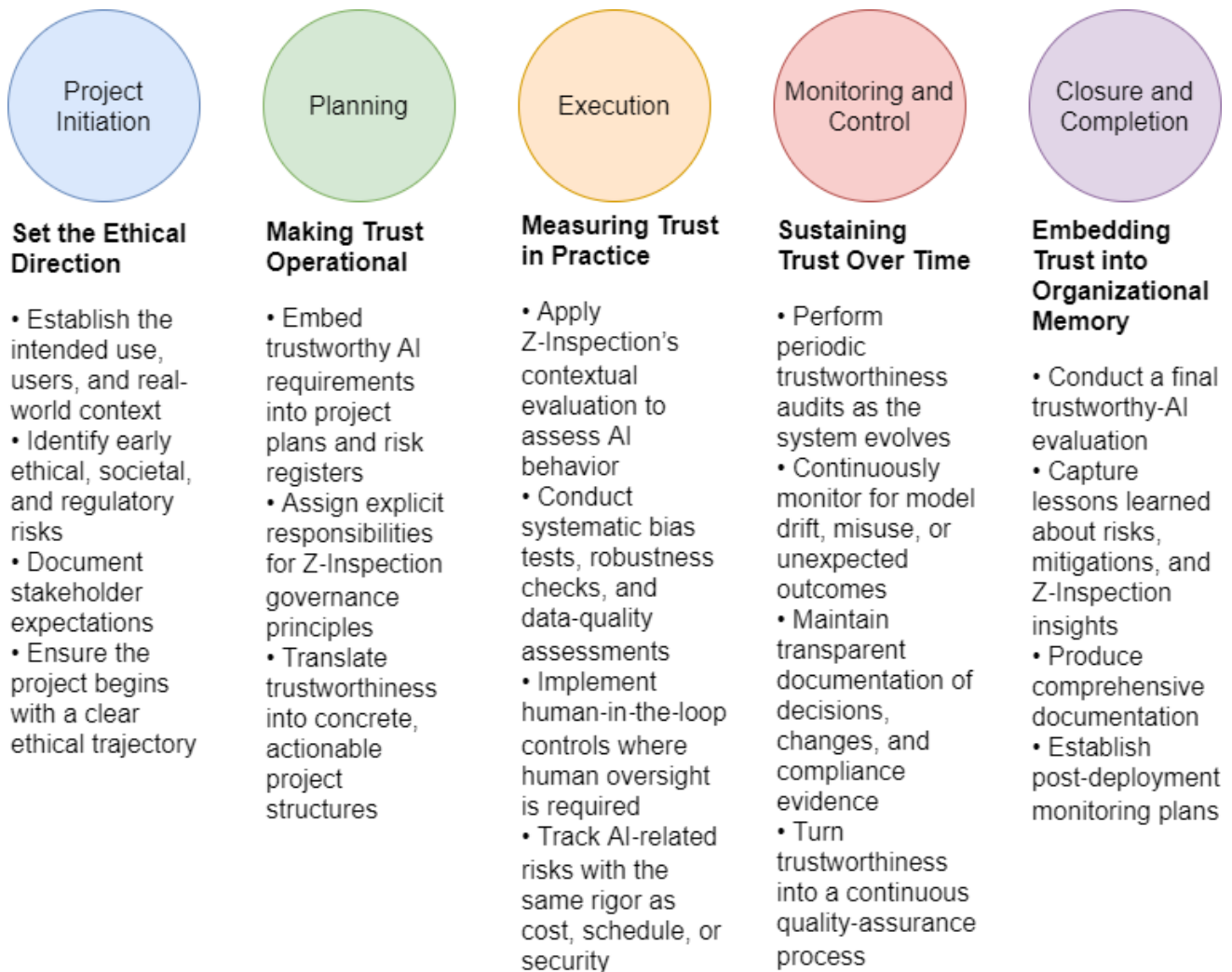


Figure 8: Diagram of the project phases using the trustworthy AI perspective proposed.

Z-Inspection® emphasizes the collection of evidence, i.e., technical documentation, data-governance records, user studies, risk analyses, and more. In project management context, this evidence becomes part of the project's documentation baseline. It also supports traceability and auditability. Furthermore, the evidence informs risk logs, change requests, and quality-management plans. All this strengthens the project's ability to demonstrate accountability and compliance.

It is important to note that Z-Inspection® is not a one-time audit but a reflexive and iterative process. When integrated into project management, Z-Inspection® methodology encourages continuous reassessment as the system evolves. It also supports agile and hybrid project models when applied. Iterative and reflective management ensures that ethical and societal considerations adapt to new findings, data, or stakeholder feedback. Embedding Z-Inspection® into project management contributes to institutionalizing trustworthy AI practices and at the same time creating repeatable governance workflows. This enhances organizational readiness for regulatory frameworks (such as the EU AI Act). It also can strengthen public trust and internal accountability.

As a summary, integrating Z-Inspection® into project-management processes creates a structured, evidence-based, and multidisciplinary approach to trustworthy AI. It transforms ethical assessment into a continuous project activity aligned with lifecycle phases, governance structures, and decision-making processes. This integration supports both regulatory compliance and the broader goal of innovating and developing AI systems that are socially responsible, transparent, and aligned with human values.

Appendix 3 contains a practical checklist for integrating the Z-Inspection® method and ethically sustainable project management into its lifecycle stages, governance and decision-making processes, in line with the AI HLEG principles (ALTAI, 2020) and the EU AI Act (2026).

6.2.7 Strategic Benefits

Integrating AI trustworthiness assessment into project management offers several strategic benefits. It enhances risk management by addressing AI-specific uncertainties, strengthens stakeholder trust by demonstrating ethical diligence, and supports regula-

tory compliance through structured documentation and evaluation. More broadly, it shifts AI governance from a reactive and compliance-driven activity to a proactive component of project strategy (Hillson, Chapter 6, 2009). In doing so, it positions project managers as key actors in ensuring that AI technologies contribute positively to healthcare innovation while safeguarding societal and ethical values.

It is important to note that the biggest challenge is not the framework itself, but the mindset. Project teams need to see trustworthy AI instead of friction:

- A way to reduce long-term risks in projects.
- A way to improve product quality in innovation and development.
- A way to build customers and users' trust.
- A competitive advantage for companies and organizations.

When PMs champion all of this, it can change the entire dynamics of projects in AI-based innovation and development. In general, we are moving towards a world where trusted AI checkpoints are standard as security audits and AI risk registers sit alongside technical and business risk registers. As a result, project managers become orchestrators of ethical, safe, and trustworthy AI innovation and development – key players in the execution of organizational strategies and success.

6.3 Analytical Reflection on the Framework Supporting the Results

The following reflections were obtained by analyzing *Microsoft 365 Copilot* prompt engineering (Ye et al., 2024) responses to our eight thematic questions, as shown in **Appendix 2**. Through the thematic areas explored in the interview, a coherent picture emerges of an industry navigating both the opportunities and the uncertainties introduced by AI. The insights gathered reveal a shared recognition that AI is becoming increasingly embedded in project workflows, yet its adoption remains contingent on the ability of organizations to manage ethical, technical, and regulatory risks in a systematic and transparent manner.

Our reflection based on **Appendix 2** underscores the multidisciplinary and highly regulated nature of healthcare product development. Project managers, engineers, regulatory specialists, and other stakeholders are already encountering AI in various forms:

from predictive analytics and automation tools to AI components embedded within medical products. This exposure shapes a pragmatic understanding of AI's potential to enhance efficiency, improve forecasting accuracy, and support complex decision-making. At the same time, however, it highlights the limitations of current tools, particularly regarding explainability, data quality, and integration with established quality management systems.

Across our findings based on **Appendix 2**, the concept of trustworthy AI emerges as a central concern. It consistently emphasizes the importance of transparency, fairness, human oversight, and robust data governance. These principles are not viewed as abstract ideals but as operational requirements that directly influence the safety, trustworthy, and regulatory acceptability of AI-driven systems. AI trustworthiness concerns have already influenced decisions about AI adoption, which demonstrates that ethical and governance considerations are embedded in day-to-day project realities. Our reflections from **Appendix 2** also identify a range of socio-technical risks, which include algorithmic bias, over-reliance on automated outputs, and ambiguity in accountability. This extends beyond technical performance and requires holistic evaluation.

A key finding of this thesis is the strong support for integrating AI trustworthiness assessment into project management processes (**Appendix 2**). The assessments should not be isolated activities but should be embedded throughout the project lifecycle from early feasibility analysis to post-deployment monitoring. This integration is seen as essential for aligning AI development with regulatory expectations, particularly considering the EU AI Act. However, our reflections also acknowledge practical challenges, including unclear ownership of AI governance tasks, resource constraints, and the need to adapt existing project management and quality systems to accommodate new assessment methods.

The Z-Inspection® method is viewed as a promising approach for addressing these challenges, as noted in **Appendix 2**. Its emphasis on contextual inquiry, stakeholder engagement, and socio-technical evaluation aligns well with the complexities of healthcare product development. The analysis recognizes its potential to strengthen compliance, enhance risk management, and support informed decision-making. At the same time, it

needs to be noted that the Z-Inspection® method may require adaptation to fit organizational workflows and may be resource-intensive in its current form. These observations suggest that while Z-Inspection® offers a valuable foundation, its practical implementation will depend on organizational readiness and the availability of appropriate competencies.

Our outlook reflects both optimism and caution. AI is expected to play an increasingly significant role in project management, particularly in areas such as predictive analytics, documentation automation, and risk forecasting. However, responsible and effective use of AI will require organizations to invest in governance structures, training, and data quality. Our reflections based on **Appendix 2** anticipate that regulatory expectations will continue to evolve and make proactive preparation essential. The need for cultural change is also highlighted, as teams adapt to new ways of working and develop a shared understanding of what trustworthy AI means in practice.

To summarize our reflections, the integration of trustworthy AI into healthcare project management is both necessary and achievable, but it requires deliberate effort, structured processes, and cross-functional collaboration. Our insights provide a strong foundation for developing a framework that supports the systematic assessment of AI trustworthiness within project management contexts.

The responses from the chatbot conversation that we obtained as the basis for the above discussion seem to be in line with other findings in this thesis. Based on the experiences of the FEHLS project and an extensive literature review, we have no concerns about, for example, AI hallucinations (Athaluri et al., 2023). We believe that using the chatbot responses enriches our insights into the proposed framework.

At this point, we can conclude that the results of this thesis answer research questions **RQ2** and **RQ3**. That is, our second hypothesis **RH2**:

“Integrating the Z-Inspection® methodology into project management processes enhances the effectiveness of AI project governance by introducing structured decision points, clearer accountability, and more comprehensive risk-mitigation practices.”

answers **RQ2** with a testable statement. Similarly, our third hypothesis **RH3**:

“Alignment of the Z-Inspection® methodology with EU legislative requirements facilitates greater regulatory compliance and reduces uncertainty regarding legal liability and organizational responsibility in AI-driven projects.”

answers **RQ3** with a demonstrable claim.

7 Discussion

The purpose of this discussion is to interpret the findings and reflections presented in earlier sections considering existing academic literature on trustworthy artificial intelligence, project management, and healthcare product development and innovation. By examining how the perspectives of practitioners align with, extend, or challenge established theoretical frameworks, this discussion highlights the broader implications of integrating AI trustworthiness assessment, particularly through the Z-Inspection® method, into project management practices.

Overall, this thesis' findings align closely with the existing literature on trustworthy AI, AI governance, and project management. They reinforce the view that AI introduces both opportunities and risks, and that managing these effectively requires structured, context-sensitive approaches. Our findings extend the literature by highlighting the practical challenges and organizational realities of integrating trustworthy AI assessment into project management, particularly in the highly regulated context of healthcare product innovation and development, as the EU AI Act requires.

The insights gained in this thesis underscore the need for frameworks like Z-Inspection® that bridge the gap between ethical principles and operational practice. They also highlight the central role of project managers as coordinators of multidisciplinary efforts to ensure that AI systems are safe, effective, and aligned with societal values. In our opinion, these conclusions provide a strong foundation for integrating trustworthy AI assessment into the project management lifecycle.

We acknowledge that further research in this emerging field is needed to better understand, for example, practical outcomes and their implications for organizational governance and management.

7.1 Liability and Regulatory Aspects

More routine legal challenges related to artificial intelligence will almost certainly soon become more common, as Scherer (2015) wrote already in 2015. Questions such as who, or actually what, should be held responsible when, for example, an autonomous vehicle

causes an accident (Abdel-Aty & Ding, 2024), a plane crashed due to a malfunction in artificial intelligence-controlled technology (The House Committee on Transportation and Infrastructure, 2020), or how far physicians can rely on intelligent diagnostic tools without increasing their exposure to malpractice claims (Chew et al., 2025), will arise with growing frequency as “*smart*” technologies spread across an expanding range of industries.

As an example, let us consider the fatal crashes of Boeing 737 MAX aircraft in October 2018 and March 2019, which led to tragic deaths and the global grounding of the Boeing 737 MAX fleet (Boeing 737 MAX, 2026). In addition to these tragic accidents and the loss of hundreds of lives, the unfortunate crashes have had a huge impact on Boeing’s and its customers’ businesses, causing huge financial losses and reputational damage. It makes intuitively sense to ask: Who is to blame – or is there such a thing as blame? Could the flight crew, other personnel, Boeing engineers, software or AI developers, or Boeing’s executives and managers be to blame? Hardly, albeit lawyers may often demand large compensation from the parties involved if they see victims in such cases. Many of Boeing’s executives and senior employees were forced to resign after the accidents (Boeing 737 MAX, 2026). In the accident investigation, it was discovered that the plane crashes were caused by a poorly designed flight control system - a single faulty sensor in the Maneuvering Characteristics Augmentation System or MCAS. After the investigations, Boeing promised to fix the planes’ MCAS software, and the company had to acknowledge that their weak regulatory environment in the development surroundings of Boeing’s MCAS (AI-driven) systems was one of the factors contributing to the accidents (Boeing 737 MAX, 2026). We could rightly ask in light of the thesis topic: Could a method like the Z-Inspection® in managerial processes of product development teams have prevented this unfortunate event? In hindsight, we cannot know.

Many of the legal issues in AI-product development have received surprisingly little attention from legal scholarships, as Scherer (2015) notes. Traditional regulatory tools, such as product licensing, oversight of research and development, and tort liability or compensation for damages, are poorly suited to the risks posed by intelligent and autonomous systems. *Ex ante* regulation is difficult because AI development can be dis-

creet or indistinguishable (requiring minimal physical infrastructure), discrete (with components designed independently), diffuse (involving contributors scattered across the globe), and opaque (making it hard for outsiders to detect potentially harmful features), as Scherer (2015) ably notices. Meanwhile, the autonomous nature of AI complicates foreseeability and control, potentially undermining the effectiveness of *ex post* regulation, especially when catastrophic risks are involved. These challenges are compounded by the persistent difficulty of defining what “*artificial intelligence*” covers in the end, as illustrated in **Figure 9**.

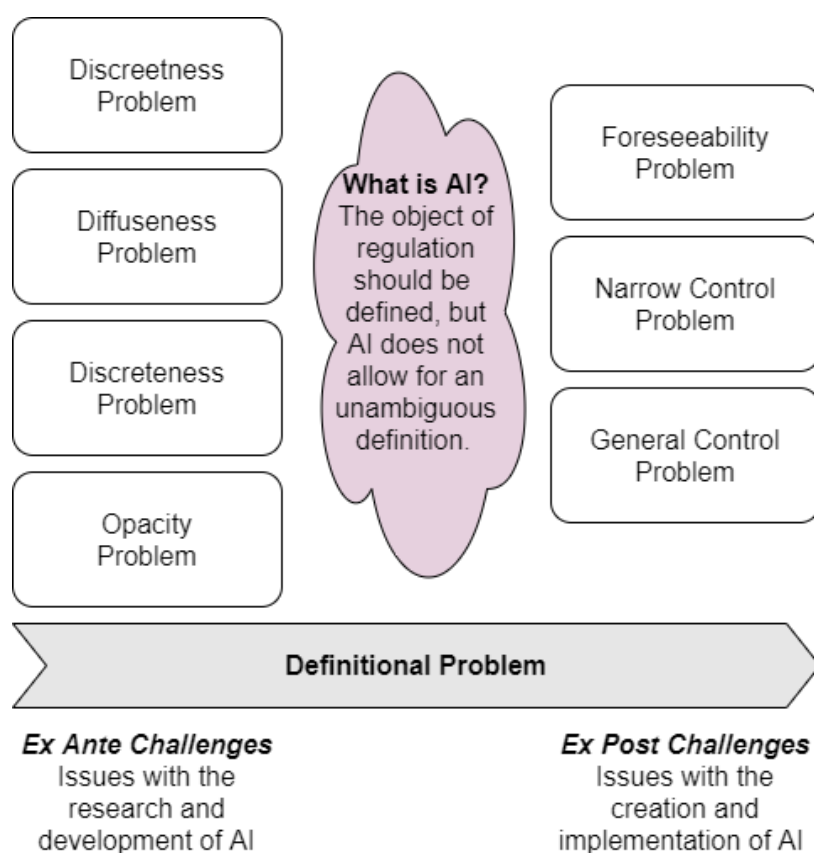


Figure 9: Regulatory challenges in artificial intelligence. Adapted from Scherer (2015).

The growing presence of AI in society and the economy creates both practical and conceptual problems for the legal system. Practically, the way AI is built and the challenge of controlling autonomous systems make regulation difficult (Scherer, 2015). Conceptually, assigning moral and legal responsibility for harm caused by autonomous machines,

and even defining AI itself, poses deep challenges, as Scherer (2015) argues. Some of these issues are unique to AI, while others resemble those associated with other advanced technologies. Taken together, they suggest that the legal system will struggle to manage AI-related harms and ensure that injured parties receive compensation.

Although legislatures derive their policymaking authority from democratic legitimacy, voters typically choose candidates based on broad platforms rather than specific policy positions, and they rarely know the details of bills (Scherer, 2015). This weakens the idea that legislation directly reflects popular will. The demands of campaigning and fundraising also limit lawmakers' time for substantive policy work, while pressure from influential interest groups may push legislators toward positions that diverge from their constituents' preferences, according to Scherer (2015).

Despite these limitations, legislative bodies remain the best suited institution to making valued policy decisions (Scherer, 2015). Scherer points out also that the main weakness in regulating emerging technologies is a lack of specialized expertise. Fortunately, the legislative bodies can compensate for this by delegating certain policymaking responsibilities. Such delegation may occur internally, through committees and subcommittees, or externally, to agencies, courts, or private bodies. Committee delegation helps narrow the expertise gap relative to administrative agencies.

7.2 Insights on Education to Support Artificial Intelligence Regulation

As Hughes et al. (2025) describe, modern project management education must incorporate ethical reasoning, awareness of algorithmic bias, and oversight capabilities to prepare future professionals for responsible AI-enabled decision-making. AI-supported learning tools, risk-assessment technologies, and experiential simulations can enrich teaching and strengthen digital fluency in project environments. To meet these demands, educators and institutions need to redesign training programs through partnerships, curriculum updates, and industry-aligned learning pathways to develop adaptable, AI-literate project managers.

Hughes et al. (2025) also emphasize that project managers must embrace principled and forward-looking leadership rooted in fairness, accountability, and stakeholder well-being

to guide the ethical integration of AI into projects. The challenges posed by AI, such as opacity, bias, and concerns over data integrity, require transparent and explainable systems supported by robust data-governance practices throughout the project lifecycle as Hughes et al. (2025) write. Leading AI-driven projects therefore calls for a shift toward distributed, ethically grounded leadership models, including ethical and servant leadership, reinforced by emerging frameworks for AI governance and the development of ethical competencies. Such a framework is proposed in this thesis.

Project managers must actively address also the environmental implications of AI and automation, ensuring that sustainability takes precedence over unrestrained technological adoption (Hughes et al., 2025). Although these technologies can improve efficiency, they also generate hidden ecological costs including increased energy use, emissions, and electronic waste that must be reduced through environmentally innovative solutions. As Hughes et al. (2025) call, the 2030 PM goals aim to train effective leadership of project managers requiring balancing technological advantages with ecological responsibility by embedding sustainability into decision-making, supply-chain practices, and lifecycle planning. Similarly, Mariani et al. (2025) reminds us that training project professionals in project management, leadership, and organization is an essential prerequisite for creating a better future. At this point, we wish to recall that the Z-Inspection® method takes also environmental and other "soft" aspects into account in the process.

It is a true and crucial issue to comprehend that trustworthy AI assessment as part of project management requires also training for engineers and developers of AI-based solutions. Trustworthy AI assessment integrated into project management practices, as proposed in this thesis, is not enough. But AI-developers, as well as AI-product development teams and managers, should understand ethical principles guided by the EU AI legislation in their innovation and development work.

Current education programs may not provide students with sufficient ethical and regulatory insights to understand the principles of EU AI legislation (Dalla et al., 2026). AI-developers need to adapt their practices and promote a culture of compliance and ethical principles in order to avoid significant fines or even market exclusion for themselves or their employers. These issues highlight the responsibility of universities and other

training providers to strengthen their programs by integrating fundamental rights, regulation, and ethics more thoroughly into the curricula of future generations of AI professionals, as Dalle et al. (2026) clarify.

7.3 Geopolitical Competitive Context

Gaske (2023) introduces a modified version of the so-called *Innovation Trilemma* (Brummer & Yadav, 2018) to guide the analysis and understanding of the external conditions that will continue to guide public policy from a high-level perspective. *Innovation Trilemma* means that if one is seeking to: (1) provide clear rules; (2) maintain market integrity; and (3) encourage innovation, regulators can at best achieve only two of these three goals.

Moreover, Finocchiaro (2024) reminds us that the EU has chosen a horizontal approach to its AI regulation framework. The inherent limitation of this kind of approach is that, because the rules are not designed to address specific sector-based issues or fill legal gaps, the regulation must be applicable across all domains, such as from healthcare to finance. This limitation means that the regulation is not specifically targeted or clearly problem-solving measures but rather it provides broad and foundational provisions intended to establish a general framework within which AI systems operate now and in the future.

So far, aside from its geopolitical significance, which is arguably regulation's core rationale, the proposed EU legislation primarily outlines an administrative structure governing the placement of AI systems on the market (Finocchiaro, 2024). This overarching framework will eventually need to be supplemented by more detailed technical standards and specifications. In this thesis, in turn, we address this very shortcoming, or even its research gap, by integrating the Z-Inspection® method into project management in accordance with the EU AI legislation, supporting innovation and development work in healthcare product development projects in accordance with it.

Besides in addition to the EU, many countries, including the United States, have openly treated AI capabilities as a strategic asset in global competition. From the perspective of private-sector rivalry, it is reasonable to assume that federal pressure to accelerate in-

novation will influence policymakers' willingness to support advanced AI development. Understanding the competitive corporate environment helps clarify how policy can encourage innovation while still addressing systemic risks and ensuring legal certainty, as Finocchiaro (2024) addresses.

Finocchiaro (2024) also emphasizes that by recognizing innovation as a central priority, policymakers may be reluctant to unfairly hinder or penalize "*first movers*" developing new AI technologies. This is especially held when major global companies are investing heavily in proprietary research, development, and startup ecosystems. For instance, *Google's* 2014 acquisition and continued backing of *DeepMind* is one such example (The Guardian, 2014), while *Microsoft* and *OpenAI's* partnership with the venture accelerator *Neo* to support emerging AI companies illustrates another (GeekWire, 2023). At the same time, national policies and corporate strategies that promote AI innovation inevitably shape systemic risks and may narrow the realistic or politically acceptable options for reducing regulatory uncertainty.

It can be stated (Finocchiaro, 2024) that regulators have constrained themselves by adopting a narrow view of competitive incentives, particularly by overlooking the role of scale in antitrust evaluations. Enforcement actions against large technology platforms have been criticized for failing to consider how data can enhance internal AI capabilities across an entire organization. A frequently cited example is the European Commission's advertising-focused restrictions on *Google's* use of biometric data from its *FitBit* acquisition (BEUC, 2020). These limits did not address whether such data might be used outside advertising. For instance, by *DeepMind* to improve AI systems deployed throughout *Google* and this way potentially strengthening the company's data-driven advantages in both scope and scale. In short, a firm that controls consumer products, like *Google*, generates vast amounts of user data, alongside powerful AI systems, continues to benefit even when regulatory constraints are narrowly drawn, as Finocchiaro (2024) recalls.

In this thesis, one of our goals is to provide an academic perspective on the integration of AI trustworthiness and project management principles, so that companies from micro to macro levels can utilize our findings in their organizations. We want to emphasize that regulations, such as the EU AI Act, will equally affect companies of all sizes that are de-

signing high-risk AI-driven applications. However, to follow the law does not necessarily require large external investments, but the principles of AI trustworthiness can be implemented also with a low threshold. The key is to understand the ethical AI principles.

7.4 Adopting Trustworthy AI in Healthcare Innovation Projects

Findings of this thesis indicate that AI is increasingly present in healthcare product development, both as a project management tool and as a component of the products themselves. This aligns with the broader trend identified in the literature, where AI is seen as a catalyst for improved efficiency, predictive accuracy, and decision support in complex environments (Topol, 2019; He et al., 2019). Our study emphasis on AI's role in forecasting, risk identification, and documentation automation. This also echoes Marnewick and Marnewick's (2020) observation that AI augments project managers' analytical capabilities rather than replacing their strategic functions.

However, the empirical data also highlight persistent challenges related to explainability, data quality, and integration with existing systems. These concerns mirror findings from Morley et al. (2020), who argue that the gap between high-level AI ethics principles and practical implementation remains significant. Our thesis' findings reinforce this by suggesting that while AI tools offer value, their adoption is constrained by technical and organizational limitations.

7.4.1 Trustworthy Artificial Intelligence as a Practical Imperative

Our findings consistently emphasize transparency, fairness, human oversight, and data governance as essential components of trustworthy AI. This aligns closely with the principles articulated by the European Commission (AI HLEG, 2020) and Floridi and Cowls (2019), who frame trustworthiness as a multidimensional construct encompassing ethical, technical, and societal considerations. This thesis suggests that these principles are not merely theoretical ideals but operational necessities in healthcare contexts, where patient safety and regulatory compliance are paramount.

This thesis reveals that trustworthiness concerns have already influenced decisions about AI adoption. The observation by Jobin et al. (2019) supports this, since organizations increasingly recognize the reputational and operational risks associated with opaque or biased AI systems. In healthcare product development, where regulatory scrutiny is high, the need for trustworthy AI becomes even more pronounced.

7.4.2 Socio-Technical Risks and Ethical Considerations

The socio-technical risks including algorithmic bias, over-reliance on automated outputs, and ambiguity in accountability are well documented in the literature. For example, Raji et al. (2020) highlight similar concerns in their work on algorithmic auditing, noting that AI systems often introduce new forms of risk that traditional governance mechanisms are ill-equipped to manage. Our findings reinforce this perspective, demonstrating that practitioners are acutely aware of the ethical and societal implications of AI, even when these implications extend beyond technical performance (or imagination).

Moreover, reflections on regulatory frameworks such as the Medical Device Regulation (MDR) and the EU AI Act align with academic discussions emphasizing the growing importance of compliance in AI governance (Benjamins et al., 2020). These regulations underscore the need for rigorous documentation, validation, and human oversight. This is also what we emphasize in the proposed framework.

7.4.3 Integrating Trustworthiness Assessment into Project Management

A key contribution of this thesis is the strong support for embedding trustworthy AI assessment throughout the project lifecycle. This perspective aligns with the project management literature, which emphasizes the importance of integrating risk management, quality assurance, and stakeholder engagement into all phases of a project (PMI, Chapter 3, 2021; Turner, 2009). Our findings suggest that trustworthy AI assessment should be treated similarly, that is, as a continuous, cross-functional activity rather than a discrete task.

This finding resonates with Morley et al. (2020), who argue that operationalizing AI ethics requires structured tools and processes that translate principles into practice. Our results indicate that practitioners see value in such tools but also recognize the challenges of integrating them into existing workflows. These challenges include unclear ownership, resource constraints, and the need for new competencies. These are also issues that the literature identifies as common barriers to AI governance maturity (Zicari et al., 2021).

7.4.4 The Role of the Z-Inspection® Method

The Z-Inspection® method emerges from the findings as a promising approach for addressing the socio-technical complexities of AI in healthcare (Kringen et al., 2026). Emphasis on contextual inquiry, stakeholder involvement, and holistic evaluation features that align with the description by Zicari et al. (2021 & 2022) of the method as a comprehensive framework for assessing trustworthy AI. The method's alignment with regulatory expectations further strengthens its relevance in healthcare product development. However, the empirical data also highlight practical limitations, such as the Z-Inspection® method's resource intensity and the need for adaptation to organizational workflows. These concerns echo broader critiques in the literature regarding the scalability and operational feasibility of ethical AI assessment frameworks (Morley et al., 2020). The findings suggest that while Z-Inspection® offers a strong conceptual foundation, its successful implementation will depend on organizational readiness and the availability of appropriate expertise. Our additional research hypothesis (**RH4**) supports this view:

“Organizations adopting a Z-Inspection®-integrated project management framework will demonstrate higher levels of organizational readiness for trustworthy AI implementation, particularly in socio-technical and ethical dimensions.”

7.4.5 Organizational Readiness and Future Outlook

Organizational readiness reveals a landscape characterized by growing awareness but limited formalization of AI governance practices. This aligns with the observation of Jobin

et al. (2019) that many organizations are still in the early stages of developing systematic approaches to AI ethics. The need for cross-functional competencies spanning technical, regulatory, and ethical domains further supports the literature's emphasis on multidisciplinary collaboration as a cornerstone of effective AI governance (Floridi & Cowls, 2019). Looking ahead, the results of this thesis anticipate that AI will become increasingly integrated into project management and product innovation and development, if organizations invest in governance structures, training, and data quality. This future-oriented perspective aligns with broader academic discussions about the evolving role of AI in healthcare and the need for proactive preparation in response to emerging regulatory frameworks (Benjamins et al., 2020). Our fifth additional research hypothesis (RH5) is in line with the above observation:

“The integration of the Z-Inspection® methodology into project management practices supports better stakeholder understanding and acceptance of trustworthy AI requirements, especially in regulated sectors such as healthcare.”

and suggests that further research is needed on this topic.

7.5 Implications for Practice

The findings of this study highlight several practical implications for organizations seeking to integrate artificial intelligence into healthcare product development while ensuring that such AI systems remain trustworthy, compliant, and aligned with ethical and societal expectations. These implications address project management practices, organizational structures, competencies, and governance mechanisms that are essential for responsible AI adoption.

A central implication is the need to integrate trustworthy AI assessment into all phases of the project lifecycle rather than treating it as a discrete or optional activity. Project initiation should include early feasibility assessments that consider ethical, regulatory, and data-related risks. Planning should incorporate trustworthiness criteria into risk registers, quality plans, and stakeholder engagement strategies. Execution and monitoring phases should include continuous evaluation of model behavior, data integrity, and human oversight mechanisms. Embedding these activities into established project man-

agement processes ensures that trustworthiness becomes a routine part of project governance rather than an afterthought.

The socio-technical nature of AI requires collaboration across technical, regulatory, clinical, and managerial domains. Organizations should establish cross-functional AI governance teams or committees responsible for overseeing trustworthiness assessments, coordinating documentation, and ensuring alignment with regulatory expectations. These structures can help clarify ownership of AI-related tasks, reduce ambiguity in accountability, and support consistent decision-making across projects. Clear governance mechanisms also facilitate the integration of frameworks such as Z-Inspection[®], which rely on multidisciplinary perspectives.

Our findings indicate that organizations often lack the necessary competencies to evaluate AI trustworthiness comprehensively. To address this gap, organizations including public educators such as universities, should invest in training programs that build literacy in AI ethics, data governance, explainability, and regulatory requirements. Project managers, in particular, need the ability to interpret AI outputs, challenge automated recommendations, and coordinate trustworthiness assessments. Developing these competencies will enable teams to engage more effectively with AI systems and reduce the risk of over-reliance or misinterpretation.

Given the increasing regulatory scrutiny surrounding AI, particularly in the EU with the emergence of the EU AI Act, organizations must ensure that their governance practices align with evolving legal requirements. This includes maintaining robust documentation, ensuring traceability of decisions, validating model performance under real-world conditions, and implementing human oversight mechanisms. Integrating trustworthiness assessment into quality management systems can help organizations demonstrate compliance and reduce regulatory risk. Frameworks such as Z-Inspection[®] can serve as valuable tools for structuring these activities, provided they are adapted to organizational workflows.

Data quality emerged as a critical determinant of AI safety and trustworthiness. Organizations should prioritize investments in data governance, including processes for data collection, cleaning, labeling, and monitoring. High-quality data not only improves fed-

erated machine learning models performance but also supports transparency, explainability, and regulatory compliance. Additionally, organizations should ensure that their technical infrastructure can support the integration of AI tools into existing project management systems without compromising security or interoperability.

Cultural readiness is essential for the successful adoption of trustworthy AI. Organizations should promote a culture that values critical engagement with AI tools, encourages questioning of automated outputs, and recognizes the importance of ethical considerations. This includes addressing concerns about job displacement, clarifying the role of AI as an augmentative rather than replacement technology, and fostering open dialogue about risks and responsibilities. A culture of responsible AI use can enhance trust among project teams and support more effective integration of AI into daily workflows.

The Z-Inspection® method offers a structured approach to assessing trustworthy AI, but its practical implementation requires adaptation to organizational contexts. Organizations should tailor the method to fit their project management processes, resource constraints, and regulatory obligations. This may involve developing simplified checklists, templates (such as in **Appendix 3**), or decision-support tools that capture the core principles of Z-Inspection® while remaining feasible for routine use. Operationalizing such frameworks can help organizations move from high-level ethical principles to actionable practices.

Finally, organizations should anticipate that AI will become increasingly embedded in both project management and product development. Preparing for this future requires proactive investment in governance, competencies, and infrastructure. Organizations that develop mature AI governance capabilities will be better positioned to leverage AI's benefits while mitigating risks and meeting regulatory expectations. Early preparation will also reduce the burden of compliance as regulatory frameworks become more stringent.

7.6 Implications for Policy

The findings of this study highlight several important implications for policymakers, regulators, and standard-setting bodies responsible for governing the development and use

of artificial intelligence in healthcare product innovation and development. As AI becomes increasingly embedded in both project management processes and healthcare technologies, policy frameworks must evolve to ensure that innovation proceeds in a safe, transparent, and ethically grounded manner.

According to our findings, especially at the SME level, it consistently emphasized the need for clearer, more actionable regulatory guidance on how to operationalize trustworthiness principles such as transparency, explainability, fairness, and human oversight. While frameworks like the EU AI Act and MDR articulate high-level expectations, organizations often struggle to translate these into concrete processes. Policymakers should therefore develop more detailed guidelines, templates and sector-specific examples that help organizations implement trustworthiness assessments throughout the project lifecycle.

This study suggests that socio-technical evaluation methods, such as Z-Inspection[®], are well suited to the complexities of healthcare AI. Policymakers could encourage or mandate the use of such frameworks by integrating them into regulatory guidance, conformity assessment procedures, or quality management standards. Doing so would help ensure that ethical and societal risks are systematically evaluated alongside technical performance.

A recurring theme in our findings is the lack of organizational competencies needed to evaluate AI trustworthiness effectively. Policymakers can address this gap by funding training programs, developing competency frameworks, and supporting cross-disciplinary education initiatives that bring together expertise in AI, ethics, regulation, and project management. Such initiatives would help build a workforce capable of navigating the socio-technical challenges of AI in healthcare.

Challenges related to unclear accountability and limited transparency from AI vendors should force policymakers to strengthen general requirements for documentation, AI model explainability, and disclosure of training data characteristics. Clearer accountability structures, particularly for high-risk AI systems, would help ensure that responsibilities are well defined across developers, deployers, and end-users.

7.7 Implications for Research

Our study also identifies several avenues for future research that can deepen understanding of trustworthy AI in healthcare project management and support the development of more effective governance frameworks.

While this study highlights the importance of integrating trustworthy AI assessment into project management, further research is needed to explore how this integration can be operationalized in practice. Future studies could examine specific tools, workflows, and decision-making processes that enable project managers to evaluate AI trustworthiness effectively.

Although the Z-Inspection® approach is viewed generally positively, empirical research is needed to assess its effectiveness in real-world project environments that differ from the healthcare sector. Comparative studies could examine how different trustworthiness assessment frameworks perform in terms of usability, completeness, and impact on project outcomes. Such research would help refine these frameworks and guide their adaptation to organizational contexts.

Our findings suggest that organizations vary widely in their readiness to adopt trustworthy AI practices. Future research could develop and validate maturity models that help organizations assess their current capabilities and identify areas for improvement. Such models could integrate dimensions such as management structures, competencies, culture, and technical infrastructure.

Given the multidisciplinary nature of AI in healthcare, further research is needed to understand how cross-functional teams collaborate to assess and manage AI risks. Studies could explore communication patterns, role definitions, and decision-making dynamics within AI governance teams, providing insights into how collaboration can be optimized. As the EU AI Act and other regulatory frameworks come into force, longitudinal research will be essential to understand how these policies influence organizational behavior, project management practices, and AI adoption. Such research could identify unintended consequences, compliance challenges, and best practices for navigating regulatory change.

The findings of this thesis raise some concerns about over-reliance on AI and ambiguity in accountability. Future research could examine how project teams interact with AI tools, how trust is formed or eroded, and how human oversight can be designed to support responsible decision-making. This line of inquiry would contribute to both AI ethics and project management scholarship.

From an AI-based technical perspective, an interesting future research direction could be, for example, how the large and intensive human resources used in Z-Inspection® multidisciplinary assessment teams could be replaced by automated AI-based solutions. This could make the assessment method truly accessible to small and medium-sized enterprises (SMEs) thanks to lower cost structures. Moreover, this could provide SMEs with instruments to truly comply with EU legislation at reasonable costs.

7.8 Study Limitations

While this study provides valuable insights into the integration of trustworthy artificial intelligence within project management for healthcare product innovation and development, several limitations must be acknowledged. These limitations relate to the study's methodological choices, scope, and generalizability, and they should be considered when interpreting the findings and their implications:

Limited Sample Size and Scope of Participants: The empirical components of this study were based on the experiences of the FEHLS project (Kringen et al., 2026) and partly on the prompt engineering responses (**Appendix 2**). These components reflect the specialized AI-driven nature of the topic and the limited pool of professionals with direct experience in AI, healthcare product innovation, and project management. Although the FEHLS project team members offered rich and diverse perspectives, the sample size restricts the extent to which the findings can be generalized across the broader healthcare or AI industries. Additionally, most of our findings were drawn from similar academic organizational contexts, which may have influenced the themes that emerged.

Potential Bias in Participant Perspectives: The study relies on limited self-reported experiences and perceptions, which may be influenced by individual bi-

ases, organizational culture, or personal familiarity with AI technologies. Experts who are more engaged with AI or more aware of ethical and regulatory issues may have provided more detailed or critical reflections than those with limited exposure, such as the author of this thesis. As a result, our findings may reflect a partial view of industry practices rather than a comprehensive representation.

Focus on a Specific Regulatory and Sectoral Context: The research is situated within the context of healthcare product innovation and development, a domain characterized by stringent regulatory requirements and high ethical stakes. While this focus strengthens the relevance of the findings for healthcare organizations, it limits their applicability to other sectors where AI adoption may follow different patterns or face different constraints. Similarly, the study draws heavily on European regulatory frameworks, particularly the EU AI Act, which may not fully reflect global regulatory diversity.

Conceptual Emphasis on the Z-Inspection® Method: The study centers on the Z-Inspection® method as a framework for assessing trustworthy AI. While this method is well suited to socio-technical evaluation, its prominence in the research may have shaped our reflections and the interpretation of findings. Other frameworks, such as algorithmic auditing, ISO or IEEE standards, or internal governance models, were not examined in comparable depth. As a result, the study may not capture the full range of approaches organizations could adopt to assess AI trustworthiness.

Evolving Nature of AI Technologies and Regulations: AI technologies, governance practices, and regulatory frameworks are evolving rapidly. The findings of this study reflect the state of the field at a particular moment in time and may not fully account for future developments, such as new regulatory requirements, emerging ethical guidelines, or advances in explainability and validation techniques. This temporal limitation means that some conclusions may require re-evaluation as the landscape continues to change.

Interpretive Nature of Qualitative Analysis: A part of the study employs qualitative methods, which rely on the author's interpretation of acquired information

and data. Although systematic literature review and thematic analysis were used to enhance rigor, the interpretive nature of qualitative research introduces the possibility of research bias. Different approaches might emphasize different themes or interpret participant statements in alternative ways. This limitation is inherent to qualitative inquiry but should be acknowledged when considering the study's conclusions.

As a summary of the discussion, this thesis provides a foundation for integrating trustworthy AI assessment and project management processes into a unified framework. We fully acknowledge that further research is needed to explore the potential of our approach, for instance, in other industries and to increase the practical understanding of the proposed framework.

8 Conclusion

This thesis examines how trustworthiness of artificial intelligence can be systematically assessed and integrated into project management practices within the context of healthcare product innovation and development. As AI becomes increasingly embedded in both the processes and outputs of healthcare innovation, the need for structured, transparent, and ethically grounded governance mechanisms has become more urgent. Our study addressed this need by exploring the applicability of the Z-Inspection® method as a framework for evaluating AI trustworthiness and by investigating how such assessments can be incorporated into the project management lifecycle.

Our literature review establishes that AI adoption in healthcare is accelerating, driven by its potential to enhance decision-making, improve efficiency, and support complex analytical tasks. At the same time, our review highlights significant challenges related to transparency, fairness, accountability, and regulatory compliance. These challenges are particularly acute in healthcare due to its ethical sensitivity and stringent oversight. Existing frameworks for trustworthy AI, including the EU AI Act and ethical guidelines from international bodies, provide high-level principles but often lack operational specificity. This gap underscores the need for practical trustworthy AI assessment methods that can be embedded into organizational processes.

Our empirical findings of this study reinforce these observations. The findings reveal a nuanced understanding of both the opportunities and risks associated with AI. The value of AI tools in project management forecasting, risk identification, and documentation automation is unquestionably recognized, but they also need to be highlighted in terms of ethical, data quality, explainability and integration limitations with existing innovation and development systems. Trustworthy AI emerges as a central concern with transparency, human oversight, sustainability, and fairness identified as essential prerequisites for responsible AI use under the EU AI legislation. Importantly, trustworthiness is not seen as an abstract ethical ideal but as a practical requirement for regulatory compliance, patient safety, and organizational credibility.

A key contribution of this thesis lies in demonstrating that trustworthy AI assessment should be embedded throughout the project management lifecycle. Our findings con-

sistently support the claim that evaluating AI trustworthiness must begin at project initiation, continue through planning and execution, and extend into post-deployment monitoring. This aligns with established project management principles that emphasize continuous risk management, quality assurance, and stakeholder engagement. Our findings also suggest that integrating trustworthy AI assessment into project workflows enhances decision-making, reduces uncertainty, and supports alignment with regulatory expectations.

The Z-Inspection[®] method is regarded as a promising framework for operationalizing trustworthy AI assessment. The method's emphasis on contextual inquiry, socio-technical evaluation, and stakeholder involvement aligns well with the complexities of healthcare product innovation and development. Our findings support the potential of Z-Inspection[®] methodology to strengthen compliance, improve risk management, and support transparent decision-making. However, we also recognize practical challenges, including the method's human resource intensity and the need for adaptation to organizational workflows. These insights suggest that while Z-Inspection[®] provides a strong conceptual foundation, its successful implementation requires organizational readiness, cross-functional collaboration, and tailored integration into existing quality management systems.

The implications of this thesis extend beyond individual projects. For practitioners, the findings highlight the importance of developing competencies in AI governance, fostering a culture of responsible AI use, and establishing cross-functional structures to support socio-technical evaluation. For policymakers, the study underscores the need for clearer regulatory guidance, harmonization of AI-specific and healthcare-specific requirements, and support for capacity building across the industry. For researchers, our findings point to opportunities for further investigation into operationalizing trustworthy AI assessment, evaluating the effectiveness of frameworks like Z-Inspection[®], and exploring organizational maturity in AI governance.

In conclusion, this thesis demonstrates that integrating trustworthy AI assessment into project management is both necessary and feasible. As AI continues to shape the future of healthcare product innovation and development, organizations must adopt structured,

context-sensitive approaches to ensure that these technologies are safe, transparent, and aligned with societal values. By bridging the gap between ethical principles and practical implementation, this research contributes to the development of governance practices that support responsible innovation. The proposed integration of Z-Inspection® into the project management lifecycle offers a pathway for organizations to navigate the complexities of AI adoption while maintaining the highest standards of trustworthiness and regulatory compliance.

References

- Abdel-Aty, M., Ding, S. (2024). A matched case-control analysis of autonomous vs human-driven vehicle accidents. *Nat Commun* **15**, 4931. <https://doi.org/10.1038/s41467-024-48526-4>
- AI HLEG. (2020). Ethics guidelines for trustworthy AI by Ala-Pietilä, P., Bonnet, Y., Bergmann, U., Bielikova, M., Bonefeld-Dahl, C., Bauer, W., ... & Van Wynsberghe, A. *European Commission*. Retrieved 22.2.26 from <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- AISC. (2024). IEEE Guide for Framework for Trustworthy Federated Machine Learning: IEEE Std 3187™-2024. *Artificial Intelligence Standards Committee*. <https://doi.org/10.1109/IEEESTD.2024.10807155>
- Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., ... & Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, 96, 156-191. <https://doi.org/10.1016/j.inffus.2023.03.008>
- Alevizos, V., Georgousis, I., Simasiku, A., Karypidou, S., & Messinis, A. (2023). Evaluating the Inclusiveness of Artificial Intelligence Software in Enhancing Project Management Efficiency - A Review. *arXiv preprint arXiv:2311.11159*. <https://doi.org/10.48550/arXiv.2311.11159>
- Ali, M., Naeem, F., Tariq, M., & Kaddoum, G. (2022). Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey. *IEEE journal of biomedical and health informatics*, 27(2), 778-789. <https://doi.org/10.1109/JBHI.2022.3181823>
- ALTAI. (2020). The assessment list for trustworthy artificial intelligence (ALTAI) by Ala-Pietilä, P., Bonnet, Y., Bergmann, U., Bielikova, M., Bonefeld-Dahl, C., Bauer, W., ... & Van Wynsberghe, A. *European Commission*. Retrieved 22.2.26 from <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>
- Alzubaidi, L., Al-Sabaawi, A., Bai, J., Dukhan, A., Alkenani, A. H., Al-Asadi, A., ... & Gu, Y. (2023). Towards Risk-Free Trustworthy Artificial Intelligence: Significance and

- Requirements. *International Journal of Intelligent Systems*, 2023(1), 4459198. <https://doi.org/10.1155/2023/4459198>
- Asilomar AI. (2017). The Asilomar AI Principles. *The Future of Life Institute*. Retrieved 23.2.2026 from <https://futureoflife.org/open-letter/ai-principles/>
- Athaluri, S. A., Manthena, S. V., Kesapragada, V. K. M., Yarlagadda, V., Dave, T., & Dudumpudi, R. T. S. (2023). Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*, 15(4). <https://doi.org/10.7759/cureus.37432>
- Audit Algorithms. (2020). Auditing machine learning algorithms. *A white paper for public auditors by the Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK*. Retrieved 10.6.2025 from <https://www.auditingalgorithms.net/index.html>
- Badhon, B., Chakraborty, R. K., Anavatti, S. G., & Vanhoucke, M. (2025). A multi-module explainable artificial intelligence framework for project risk management: enhancing transparency in decision-making. *Engineering Applications of Artificial Intelligence*, 148, 110427. <https://doi.org/10.1016/j.engappai.2025.110427>
- Bayazit, N. (2004). Investigating design: A review of forty years of design research. *Design issues*, 20(1), 16-29. <http://mediadesign-lab.com/database/THESIS/5.pdf>
- Benjamens, S., Dhunoo, P., & Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ digital medicine*, 3(1), 118. <https://doi.org/10.1038/s41746-020-00324-0>
- BEUC. (2020). Google's Fitbit takeover: EU merger control proves unable to protect consumers in the digital economy. *The European Consumer Organisation (BEUC)*. Press Release. Retrieved 14.2.2026 from <https://www.beuc.eu/press-releases/googles-fitbit-takeover-eu-merger-control-proves-unable-protect-consumers-digital>
- Boeing 737 MAX. (2026). Reactions to the Boeing 737 MAX groundings. *Wikipedia*. Retrieved 14.2.2026 from https://en.wikipedia.org/wiki/Reactions_to_the_Boeing_737_MAX_groundings
- Boiano, A., Di Gennaro, M., Barbieri, L., Carminati, M., Nicoli, M., Redondi, A., ... & Serio, L. (2024, October). A secure and trustworthy network architecture for federated

- learning healthcare applications. In *2024 20th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)* (pp. 124-129). IEEE. <https://doi.org/10.1109/WiMob61911.2024.10770536>
- Brummer, C., & Yadav, Y. (2018). Fintech and the innovation trilemma. *Geo. LJ*, *107*, 235. <https://www.law.georgetown.edu/georgetown-law-journal/wp-content/uploads/sites/26/2019/02/1Fintech-and-the-Innovation-Trilemma.pdf>
- Buruk, B., Ekmekci, P. E., & Arda, B. (2020). A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Health Care and Philosophy*, *23*(3), 387-399. <https://doi.org/10.1007/s11019-020-09948-1>
- Bærøe, K., Miyata-Sturm, A., & Henden, E. (2020). How to achieve trustworthy artificial intelligence for health. *Bulletin of the World Health Organization*, *98*(4), 257. <https://doi.org/10.2471/BLT.19.237289>
- Chen, C., Liu, J., Tan, H., Li, X., Wang, K. I. K., Li, P., ... & Dou, D. (2025). Trustworthy federated learning: privacy, security, and beyond. *Knowledge and Information Systems*, *67*(3), 2321-2356. <https://doi.org/10.1007/s10115-024-02285-2>
- Chew, K., Snyder, K., & Pert, C. (2025). How Physicians Might Get in Trouble Using AI (or Not Using AI). *Missouri Medicine*, *122*(3), 169. <https://pubmed.ncbi.nlm.nih.gov/40747395/>
- Christensen, C. M., Kaufman, S. P., & Shih, W. C. (2008). Innovation killers: How financial tools destroy your capacity to do new things. *Harvard business review*, *86*(1), 98-105. PMID: 18271321.
- Cutillo, C. M., Sharma, K. R., Foschini, L., Kundu, S., Mackintosh, M., Mandl, K. D., & MI in Healthcare Workshop Working Group (2020). Machine intelligence in healthcare perspectives on trustworthiness, explainability, usability, and transparency. *NPJ digital medicine*, *3*(1), 47. <https://doi.org/10.1038/s41746-020-0254-2>
- Dalla Giovanna, V., Lewkowicz, G., & Mahloul, A. (2026). Evaluating University Curricula in Preparing Future AI Developers to Comply with the AI Act: A Case Study of Belgium. In *Compliance for Artificial Intelligence Systems: Strategies, Principles and Methods* (pp. 75-95). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-12795-2_6

- Ducange, P., Marcelloni, F., Renda, A., & Ruffini, F. (2024). Federated Learning of XAI Models in Healthcare: A Case Study on Parkinson's Disease. *Cognitive Computation*, 16(6), 3051-3076. <https://doi.org/10.1007/s12559-024-10332-x>
- Duică, M. C., Vasiciuc, C. G., & Panagoreț, D. (2024). The use of artificial intelligence in project management. *Valahian Journal of Economic Studies*, 15(1), 105-118. <https://doi.org/10.2478/vjes-2024-0009>
- EU AI Act, The. (2026). The EU Artificial Intelligence Act: Up-to-date developments and analyses of the EU AI Act. Retrieved 3.2.2026 from <https://artificialintelligenceact.eu/>
- EU AI Act: Annex I, The. (2026). Annex I: List of Union Harmonisation Legislation. Retrieved 3.2. 2026 from <https://artificialintelligenceact.eu/annex/1/>
- EU AI Act: Annex II, The. (2026). Annex III: High-Risk AI Systems Referred to in Article 6(2). Retrieved 3.2.26 from <https://artificialintelligenceact.eu/annex/3/>
- EU AI Act: Explorer, The. (2026). The AI Act Explorer. Retrieved 3.2.2026 from <https://artificialintelligenceact.eu/ai-act-explorer/>
- Felicetti, A. M., Cimino, A., Mazzoleni, A., & Ammirato, S. (2024). Artificial intelligence and project management: An empirical investigation on the appropriation of generative Chatbots by project managers. *Journal of Innovation & Knowledge*, 9(3), 100545. <https://doi.org/10.1016/j.jik.2024.100545>
- Finocchiaro, G. (2024). The regulation of artificial intelligence. *Ai & Society*, 39(4), 1961-1968. <https://doi.org/10.1007/s00146-023-01650-z>
- Floridi, L., Holweg, M., Taddeo, M., Silva, J. A., Mökander, J., & Wen, Y. (2022). *capAI-A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act*. *SSRN Electronic Journal* (2022). <https://dx.doi.org/10.2139/ssrn.4064091>
- Fridgeirsson, T. V., Ingason, H. T., Jonasson, H. I., & Jonsdottir, H. (2021). An authoritative study on the near future effect of artificial intelligence on project management knowledge areas. *Sustainability*, 13(4), 2345. <https://doi.org/10.3390/su13042345>
- Gaske, M. R. (2023). Regulation priorities for artificial intelligence foundation models. *Vand. J. Ent. & Tech. L.*, 26, 1. <https://scholarship.law.vanderbilt.edu/jetlaw/vol26/iss1/2>

- GDPR. (2016). General Data Protection Regulation – GDPR. *Regulation (EU) 2016/679*. Retrieved 4.2.2026 from <https://gdpr-info.eu/>
- GeekWire. (2023). GeekWire Podcast: Microsoft, OpenAI, and the AI opportunity for startups, with Neo CEO Ali Partovi. *GeekWire Podcasts*, Mar 25, 2023. Retrieved 14.2.2026 from <https://www.geekwire.com/2023/geekwire-podcast-openai-microsoft-and-the-ai-opportunity-for-startups-with-neo-ceo-ali-partovi/>
- Genus, A., & Stirling, A. (2018). Collingridge and the dilemma of control: Towards responsible and accountable innovation. *Research policy*, 47(1), 61-69. <https://doi.org/10.1016/j.respol.2017.09.012>
- Gil, J., Martinez Torres, J., & González-Crespo, R. (2021). The application of artificial intelligence in project management research: A review. <https://doi.org/10.9781/ijimai.2020.12.003>
- Giray, L. (2023). Prompt engineering with ChatGPT: a guide for academic writers. *Annals of biomedical engineering*, 51(12), 2629-2633. <https://doi.org/10.1007/s10439-023-03272-4>
- Guardian, The. (2014). Google buys UK artificial intelligence startup Deepmind for £400m. *The Guardian*, Jan 2014. Retrieved 14.2.2026 from <https://www.theguardian.com/technology/2014/jan/27/google-acquires-uk-artificial-intelligence-startup-deepmind>
- Guo, J., Pietzuch, P., Paverd, A., & Vaswani, K. (2024). Trustworthy AI using Confidential Federated Learning: Federated learning and confidential computing are not competing technologies. *Queue*, 22(2), 87-107. <https://doi.org/10.1145/3677390>
- Happonen A. and van Gils M. (2025). Trustworthy Impact Assessment of a Federated Machine Learning System in Healthcare Product Development. *MET Research Days 2025: Abstracts*. Tampere, Finland, 30-31 October 2025. Retrieved 3.2.2026 from https://events.tuni.fi/app/uploads/sites/1273/2025/10/antti-happonen_2025.pdf
- Happonen A., van Gils M., Danelakis A., Kumelj T., Tronvik E., Hau M., Kringen P., Ekman P., Westerlund M., Moreno Sánchez P. A., Vesteghem C., and Barlous B. C. (2025). Federated Ethical Healthcare Learning Sandbox Project. *AI Day 2025 – FCAI*. Espoo, Finland, November 13 2025. Retrieved

- 3.2.2026 from https://docs.google.com/spreadsheets/d/1QNBqRA2-RoT-WgHmgNha_oTRI_NjgxSm8Y9U8irQTvV4/edit?gid=0#gid=0 (Poster #79).
- Hasani, N., Morris, M. A., Rhamim, A., Summers, R. M., Jones, E., Siegel, E., & Saboury, B. (2022). Trustworthy artificial intelligence in medical imaging. *PET clinics*, 17(1), 1. <https://doi.org/10.1016/j.cpet.2021.09.007>
- He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature medicine*, 25(1), 30-36. <https://doi.org/10.1038/s41591-018-0307-0>
- Hickman, E., & Petrin, M. (2021). Trustworthy AI and corporate governance: the EU's ethics guidelines for trustworthy artificial intelligence from a company law perspective. *European Business Organization Law Review*, 22, 593-625. <https://doi.org/10.1007/s40804-021-00224-0>
- Hillson, D. (2009). *Managing Risk in Projects*. Gower.
- Hossain, M. Z., Hasan, L., Dewan, M. A., & Monira, N. A. (2024). The impact of artificial intelligence on project management efficiency. *International Journal of Management Information Systems and Data Science*, 1(5), 1-17. <https://doi.org/10.62304/ijmisdsv1i05.211>
- House Committee on Transportation and Infrastructure, The. (2020). After 18-Month Investigation, Chairs DeFazio and Larsen Release Final Committee Report on Boeing 737 MAX. *The House Committee on Transportation and Infrastructure*, September 16, 2020. Retrieved 14.2.2026 from <https://democrats-transportation.house.gov/news/press-releases/after-18-month-investigation-chairs-defazio-and-larsen-release-final-committee-report-on-boeing-737-max>
- Hughes, L., Mavi, R. K., Aghajani, M., Fitzpatrick, K., Gunaratnege, S. M., Shekarabi, S. A. H., ... & Dwivedi, Y. K. (2025). Impact of artificial intelligence on project management (PM): Multi-expert perspectives on advancing knowledge and driving innovation toward PM2030. *Journal of Innovation & Knowledge*, 10(5), 100772. <https://doi.org/10.1016/j.jik.2025.100772>
- Ibadildin, N., Kenzhin, Z., Yeshenkulova, G., Ismailova, R., Nurguzhina, A., Nassanbekova, S., & Kadyrova, A. (2025). Artificial intelligence in project management: A bibliometric

- analysis. *Problems and Perspectives in Management*, 23(2), 252.
[https://doi.org/10.21511/ppm.23\(2\).2025.17](https://doi.org/10.21511/ppm.23(2).2025.17)
- IEEE. (2021). IEEE CertifAIEd: The mark of AI ethics. Retrieved 10.6.2025 from
<https://engagestandards.ieee.org/ieeecertifai.html>
- Jannat, S. F., Ahmed, M. S., Rajput, S. A., & Hasan, S. (2024). AI-powered project management: Myth or reality? Analyzing the integration and impact of artificial Intelligence in contemporary project environments. *International Journal of Applied Engineering & Technology*, 6(1), 1810-1820. Retrieved 22.2.2026 from [https://www.academia.edu/download/118303271/AI Powered Project Management Myth or Reality Analyzing the Integration and Impact of.pdf](https://www.academia.edu/download/118303271/AI_Powered_Project_Management_Myth_or_Reality_Analyzing_the_Integration_and_Impact_of.pdf)
- Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)*, 55(2), 1-38.
<https://doi.org/10.1145/3491209>
- Kemmerzell, N., Schreiner, A., Khalid, H., Schalk, M., & Bordoli, L. (2025). Towards a Better Understanding of Evaluating Trustworthiness in AI Systems. *ACM Computing Surveys*, 57(9), 1-38. <https://doi.org/10.1145/3721976>
- Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.
- Kortum, H., Rebstadt, J., Bösch, T., Meier, P., & Thomas, O. (2022). Towards the operationalization of trustworthy AI: integrating the EU assessment list into a procedure model for the development and operation of AI-systems. In *INFORMATIK 2022* (pp. 283-299). Gesellschaft für Informatik, Bonn.
https://doi.org/10.18420/inf2022_26
- Kringen P., Happonen A. P., Ekman P., Vesteghem C., Barlous B., Danelakis A., van Gils M., Moreno-Sánchez P. A., and Westerlund M. (2026). Trustworthy Federated Machine Learning in Healthcare: A Sociotechnical and Risk-Based Evaluation Using the Z-Inspection® Method. Accepted Mar. 2026 to *Nordic Conference on Digital Health and Wireless Solutions 2026*, Oulu, Finland, June 16-17. To appear in *Communications in Computer and Information Science*.

- Lekadir, K., Frangi, A. F., Porras, A. R., Glocker, B., Cintas, C., Langlotz, C. P., ... & Starmans, M. P. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *bmj*, 388. <https://doi.org/10.1136/bmj-2024-081554>
- Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Computers, Environment and Urban Systems*, 96, 101845. <https://doi.org/10.1016/j.compenvurbsys.2022.101845>
- Li, H., Li, C., Wang, J., Yang, A., Ma, Z., Zhang, Z., & Hua, D. (2023). Review on security of federated learning and its application in healthcare. *Future Generation Computer Systems*, 144, 271-290. <https://doi.org/10.1016/j.future.2023.02.021>
- Liang, W., Tadesse, G. A., Ho, D., Fei-Fei, L., Zaharia, M., Zhang, C., & Zou, J. (2022). Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*, 4(8), 669-677. <https://doi.org/10.1038/s42256-022-00516-1>
- Mariani, C., Aaltonen, K., por Ingason, H., Mancini, M., & Huemann, M. (2025). The future of learning: how artificial intelligence and other new technologies revolutionize project management education and foster Project Learning Intelligence. *Project Leadership and Society*, 100199. <https://doi.org/10.1016/j.plas.2025.100199>
- Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics*, 113, 103655. <https://doi.org/10.1016/j.jbi.2020.103655>
- Marnewick, C., & Marnewick, A. (2021). Digital intelligence: A must-have for project managers. *Project Leadership and Society*, 2, 100026. <https://doi.org/10.1016/j.plas.2021.100026>
- McNeill, K. L., & Martin, D. M. (2011). Claims, evidence, and reasoning. *Science and Children*, 48(8), 52. <http://searkscience.pbworks.com/w/file/70117336/2-Claim-sevidence.pdf>
- MedTech Europe. (2025). One year of AI Act: MedTech Europe calls for coherent implementation to unlock the full potential of AI in healthcare. *MedTech Europe*. Press Release. Brussels, 1 August 2025. Retrieved 17.3.2026 from

https://www.medtecheurope.org/wp-content/uploads/2025/07/250801-ai-act-one-year-anniversary_press-release.pdf

Mesa Fernández, J. M., González Moreno, J. J., Vergara-González, E. P., & Alonso Iglesias, G. (2022). Bibliometric analysis of the application of artificial intelligence techniques to the management of innovation projects. *Applied Sciences*, 12(22), 11743.

<https://doi.org/10.3390/app122211743>

MIT AI Risk Repository, The. (2026). MIT AI Risk Initiative. *MIT FutureTech 2025*. Retrieved 16.2.2026 from <https://airisk.mit.edu/>

Montreal Declaration. (2018). Montréal Declaration For A Responsible Development Of Artificial Intelligence. *An initiative of Université de Montréal*. Retrieved 23.2.2026 from

<https://montrealdeclaration-responsibleai.com/the-declaration/>

Moraes, T. (2025). Regulatory sandboxes for trustworthy artificial intelligence—global and Latin American experiences. *International Review of Law, Computers & Technology*, 39(1), 55-74. <https://doi.org/10.1080/13600869.2024.2351674>

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, 26(4), 2141-2168.

<https://doi.org/10.1007/s11948-019-00165-5>

Müller, R., Locatelli, G., Holzmann, V., Nilsson, M., & Sagay, T. (2024). Artificial Intelligence and project management: Empirical overview, state of the art, and guidelines for future research. *Project Management Journal*, 55(1), 9-15.

<https://doi.org/10.1177/87569728231225198>

Newman, J. (2023). A taxonomy of trustworthiness for artificial intelligence. *CLTC: North Charleston, SC, USA*, 1. Retrieved 22.2.2026 from

https://cltc.berkeley.edu/wp-content/uploads/2023/12/Taxonomy_of_AI_Trustworthiness_tables.pdf

Nguyen, D. C., Pham, Q. V., Pathirana, P. N., Ding, M., Seneviratne, A., Lin, Z., ... & Hwang, W. J. (2022). Federated learning for smart healthcare: A survey. *ACM Computing Surveys (Csur)*, 55(3), 1-37. <https://doi.org/10.1145/3501296>

- Niederman, F. (2021). Project management: openings for disruption from AI and advanced analytics. *Information Technology & People*, 34(6), 1570-1599. <https://doi.org/10.1108/ITP-09-2020-0639>
- Nordic Innovation. (2026). Applied Ethical AI on Nordic Patient Records. *Nordic Innovation*. Retrieved 4.2.2026 from <https://www.nordicinnovation.org/programs/applied-ethical-ai-nordic-patient-records>
- OECD, (2025). Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449. <https://doi.org/10.1017/ilm.2020.5>
- Paparc, M., & Bodea, C. N. (2024). Organising Projects for Responsible Use of Generative Artificial Intelligence in Project Management. *Informatica Economica*, 28(3). Retrieved 22.2.2026 from <https://www.revistaie.ase.ro/content/111/01%20-%20paparc,bodea.pdf>
- Peylo, C., Slama, D., Hallensleben, S., Hauschke, A., & Hildebrandt, S. (2023). VCIO based description of systems for AI trustworthiness characterisation (2022). Retrieved 11.6.2025. Retrieved 22.2.2026 from <https://www.vde.com/resource/blob/2242194/a24b13db01773747e6b7bba4ce20ea60/vcio-based-description-of-systems-for-ai-trustworthiness-characterisationvde-spec-90012-v1-0--en--data.pdf>
- Pica, M. M. (2015). *Project Life Cycle Economics: Cost Estimation, Management and Effectiveness in Construction Projects*. Gower Publishing, Ltd.
- PMI. (2021, July). *A Guide to the Project Management Body of Knowledge (PMBOK® Guide)*—Seventh Edition and The Standard for Project Management. Project Management Institute.
- Poretschkin, M., Schmitz, A., Akila, M., Adilova, L., Becker, D., Cremers, A. B., ... & Wrobel, S. (2023). Guideline for Trustworthy Artificial Intelligence - AI Assessment Catalog. *arXiv preprint arXiv:2307.03681*. Retrieved 24.2.2026 from https://www.researchgate.net/publication/372234481_Guideline_for_Trustworthy_Artificial_Intelligence_-_AI_Assessment_Catalog
- Prasetyo, M. L., Peranginangin, R. A., Martinovic, N., Ichsan, M., & Wicaksono, H. (2024). Artificial intelligence in open innovation project management: A systematic literature review on technologies, applications, and integration requirements. *Journal of Open*

Innovation: Technology, Market, and Complexity, 100445.
<https://doi.org/10.1016/j.oiitmc.2024.100445>

Radclyffe, C., Ribeiro, M., & Wortham, R. H. (2023). The assessment list for trustworthy artificial intelligence: A review and recommendations. *Frontiers in artificial Intelligence*, 6, 1020592. <https://doi.org/10.3389/frai.2023.1020592>

Rahman, A., Hossain, M. S., Muhammad, G., Kundu, D., Debnath, T., Rahman, M., ... & Band, S. S. (2023). Federated learning-based AI approaches in smart healthcare: concepts, taxonomies, challenges and open issues. *Cluster computing*, 26(4), 2271-2311. <https://doi.org/10.1007/s10586-022-03658-4>

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33-44). <https://doi.org/10.1145/3351095.3372873>

Regulation (EU) 2024/1689. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council. *The European Union*. Retrieved 3.2.2026 from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689>

Regulation (EU) 2017/745. (2017). Regulation (EU) 2017/745 of the European Parliament and of the Council. *The European Union*. Retrieved 4.2.2026 from <https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng>

Regulation (EU) 2017/746. (2017). Regulation (EU) 2017/746 of the European Parliament and of the Council. *The European Union*. Retrieved 19.2.2026 from <https://eur-lex.europa.eu/eli/reg/2017/746/oj>

Salimimoghadam, S., Ghanbaripour, A. N., Tumpa, R. J., Kamel Rahimi, A., Golmoradi, M., Rashidian, S., & Skitmore, M. (2025). The Rise of Artificial Intelligence in Project Management: A Systematic Literature Review of Current Opportunities, Enablers, and Barriers. *Buildings*, 15(7), 1130. <https://doi.org/10.3390/buildings15071130>

Sausser, B., Verma, D., Ramirez-Marquez, J., & Gove, R. (2006, April). From TRL to SRL: The concept of systems readiness levels. In *Conference on Systems Engineering Research*,

- Los Angeles, CA, April 7th-8th 2006.
<https://api.semanticscholar.org/CorpusID:16808899>
- Scherer, M. U. (2015). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harv. JL & Tech.*, 29, 353.
<https://doi.org/10.2139/ssrn.2609777>
- Schlicker, N., Baum, K., Uhde, A., Sterz, S., Hirsch, M. C., & Langer, M. (2025). How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM). *Computers in Human Behavior*, 170, 108671.
<https://doi.org/10.1016/j.chb.2025.108671>
- Shang, G., Low, S. P., & Lim, X. Y. V. (2023). Prospects, drivers of and barriers to artificial intelligence adoption in project management. *Built Environment Project and Asset Management*, 13(5), 629-645. <https://doi.org/10.1108/BEPAM-12-2022-0195>
- Shoushtari, F., Daghighi, A., & Ghafourian, E. (2024). Application of artificial intelligence in project management. *International journal of industrial engineering and operational research*, 6(2), 49-63. <https://doi.org/10.22034/ijieor.v6i2.89>
- Stahl, B. C., Antoniou, J., Bhalla, N., Brooks, L., Jansen, P., Lindqvist, B., ... & Wright, D. (2023). A systematic review of artificial intelligence impact assessments. *Artificial Intelligence Review*, 56(11), 12799-12831.
<https://doi.org/10.1007/s10462-023-10420-8>
- Stephanie, V., Khalil, I., Atiquzzaman, M., & Yi, X. (2022). Trustworthy privacy-preserving hierarchical ensemble and federated learning in healthcare 4.0 with blockchain. *IEEE Transactions on Industrial Informatics*, 19(7), 7936-7945.
<https://doi.org/10.1109/TII.2022.3214998>
- Tabassi, E. (2023), Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST Trustworthy and Responsible AI. *National Institute of Standards and Technology*. Gaithersburg, MD, [online]. <https://doi.org/10.6028/NIST.AI.100-1>
- Taboada, I., Daneshpajouh, A., Toledo, N., & De Vass, T. (2023). Artificial intelligence enabled project management: a systematic literature review. *Applied Sciences*, 13(8), 5014. <https://doi.org/10.3390/app13085014>

- Tariq, A., Serhani, M. A., Sallabi, F. M., Barka, E. S., Qayyum, T., Khater, H. M., & Shuaib, K. A. (2024). Trustworthy federated learning: A comprehensive review, architecture, key challenges, and future research prospects. *IEEE Open Journal of the Communications Society*. <https://doi.org/10.1109/OJCOMS.2024.3438264>
- Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1), 44-56. <https://doi.org/10.1038/s41591-018-0300-7>
- UN. (2023). Independent Group of Scientists appointed by the Secretary-General, Global Sustainable Development Report 2023: Times of crisis, times of change: Science for accelerating transformations to sustainable development. *United Nations*. New York. Retrieved 22.2.2026 from https://sdgs.un.org/sites/default/files/2023-09/FINAL%20GSDR%202023-Digital%20-110923_1.pdf
- Vergara, D., del Bosque, A., Lampropoulos, G., & Fernández-Arias, P. (2025). Trends and applications of artificial intelligence in project management. *Electronics*, 14(4), 800. <https://doi.org/10.3390/electronics14040800>
- Vetter, D., Amann, J., Bruneault, F., Coffee, M., Düdder, B., Gallucci, A., ... & Z-Inspection® initiative (2022). (2023). Lessons learned from assessing trustworthy AI in practice. *Digital Society*, 2(3), 35. <https://doi.org/10.1007/s44206-023-00063-1>
- Wu, W., & Liu, S. (2023). A comprehensive review and systematic analysis of artificial intelligence regulation policies. *arXiv preprint arXiv:2307.12218*. <https://doi.org/10.48550/arXiv.2307.12218>
- Yang, Z., Shi, Y., Zhou, Y., Wang, Z., & Yang, K. (2022). Trustworthy federated learning via blockchain. *IEEE Internet of Things Journal*, 10(1), 92-109. <https://doi.org/10.1109/JIOT.2022.3201117>
- Ye, Q., Ahmed, M., Pryzant, R., & Khani, F. (2024, August). Prompt engineering a prompt engineer. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 355-385). <https://doi.org/10.18653/v1/2024.findings-acl.21>
- Zhang, Y., Zeng, D., Luo, J., Fu, X., Chen, G., Xu, Z., & King, I. (2024). A survey of trustworthy federated learning: Issues, solutions, and challenges. *ACM Transactions on Intelligent Systems and Technology*, 15(6), 1-47. <https://doi.org/10.1145/3678181>

Zicari, R. V., Amann, J., Bruneault, F., Coffee, M., Düdder, B., Hickman, E., ... & Wurth, R. (2022). How to assess trustworthy AI in practice. *arXiv preprint arXiv:2206.09887*.

Zicari, R. V., Brodersen, J., Brusseau, J., Düdder, B., Eichhorn, T., Ivanov, T., ... & Westerlund, M. (2021). Z-Inspection®: a process to assess trustworthy AI. *IEEE Transactions on Technology and Society*, 2(2), 83-97.
<https://doi.org/10.1109/TTS.2021.3066209>

Appendices

Appendix 1: Identified Claims, Evidence and Reasonings in the FEHLS Study

In this **Appendix**, we present some of the results of the FEHLS study to clarify the steps of the Z-Inspection® process. The claims, evidence, and reasonings (CERs) identified in our Z-Inspection® process are presented in **Table 2**.

According to the Z-Inspection® process, the identified CERs are further mapped to the HLEG requirements for trustworthy AI and then refined into proposed mitigation strategies, tensions, and solutions (not shown here).

In the reported FEHLS study (Kringen et al., 2026), these identified CERs were also mapped to the MIT AI Risk Repository taxonomy.

Table 2: Identified claims, evidence and reasonings (CERs) of the developed FML systems. The claims for the Norwegian and Danish systems were derived directly from information provided by the developers in the co-design studies. Adapted from Kringen et al. (2026).

No.	Claim	Evidence	Reasoning
NOR-CER 1	The system derives patient-specific outcome predictions by identifying individuals with comparable clinical characteristics.	Internal experiments show that gradient-boosting models using similarity-based representations outperform linear baselines on neurology datasets. The documentation states that “ <i>gradient-boosting models...using a similarity-based representation, outperformed linear baselines</i> ”.	These findings indicate that the similarity-based representation captures clinically relevant patterns, supporting the claim that predictions are generated by comparing patients with similar clinical profiles.
NOR-CER 2	The federated learning architecture maintains patient privacy and	System descriptions report that only encrypted gradients are exchanged and “ <i>no patient-level records leave institutional infrastructures</i> ”.	Because data remain local and only encrypted updates circulate, the federated setup can be regarded as privacy-preserving and aligned

	adheres to regulatory requirements.		with regulatory requirements.
NOR-CER 3	The system functions as a clinical decision-support tool rather than an autonomous decision-maker	Documentation notes that interpretability tools such as SHAP are used and that <i>“no autonomous decision mechanism exists”</i> .	Since the tool cannot act independently and only provides explanatory information, clinical responsibility remains with practitioners, confirming its role as decision support.
NOR-CER 4	The system’s output can directly influence the treatment decisions made for individual patients.	The model processes individual neurology records and produces specific predictions, such as expected response to amitriptyline. The file notes that <i>“outputs concern the expected response to specific treatments such as amitriptyline”</i> .	Because predictions inform medication choices, they can directly affect patient care, although missing demographic data may reduce reliability.
NOR-CER 5	The system’s output may indirectly shape the decisions and practices of healthcare professionals and other stakeholders.	SHAP-based explanations <i>“have not been validated for clinical interpretation”</i> , and governance structures are not yet implemented.	These features may shape professional decision-making, workflows, and communication, creating risks such as automation bias and misplaced confidence in technical explanations.
NOR-CER 6	The system is currently positioned at an early experimental stage, corresponding to TRL 3.	Testing is limited to internal evaluation, with <i>“no external benchmarking, and no documented pathway toward clinical certification”</i>	The absence of integration, external validation, and certification processes supports classifying the system as an early proof-of-concept at TRL 3.
NOR-CER 7	The system is developed using Python as its primary	Project descriptions state that the federated pipeline and pre-	Python’s widespread use and interoperability justify its selection and support the claim

	implementation language.	processing routines “ <i>run in Python-based environments</i> ”.	regarding implementation language.
NOR-CER 8	The system’s robustness relies chiefly on preprocessing measures applied to the underlying data.	Documentation indicates “ <i>no adversarial testing, no cross-site robustness evaluation</i> ”, with robustness efforts focused on outlier removal, normalization, and handling missing data.	These preprocessing steps function as the main mechanism for stabilizing model performance in the absence of broader robustness testing.
NOR-CER 9	The system’s design enables the possibility of independent external validation.	The documentation states that “ <i>model updates, training configurations, evaluation procedures...can be shared without exposing patient-level data</i> ”.	Because artefacts and configurations can be shared while data remain local, the system is structured to enable external validation, even though none has yet occurred.

No.	Claim	Evidence	Reasoning
DEN-CER 1	The system estimates treatment effectiveness by comparing patients with similar clinical profiles	Structured variables extracted from GP records are used to estimate medication efficacy, and outputs “ <i>inform clinicians about the likelihood of successful treatment</i> ”.	Converting patient histories into structured variables enables similarity-based estimation of treatment effects, supporting the claim.
DEN-CER 2	The federated setup preserves national data sovereignty and protects sensitive patient information.	Sensitive data remain within national infrastructures, and “ <i>no raw patient data crosses borders</i> ”.	Encryption and local data retention ensure compliance with privacy requirements across jurisdictions.
DEN-CER 3	The system contributes to clinical decision-making by providing individu-	Clinicians receive individualized estimates of treatment effectiveness, and the system “ <i>provides cli-</i>	These outputs inform prescribing decisions and help clinicians assess treatment options.

	alized predictions of treatment response.	<i>nicians with estimates of how effective a medication is likely to be</i> ".	
DEN-CER 4	The system enhances transparency through interpretable model outputs and accompanying documentation.	SHAP-based explanations " <i>show how input variables influence predictions</i> ", and documentation outlines system behavior and limitations.	These elements enhance transparency for users and evaluators by clarifying model logic and constraints
DEN-CER 5	The system employs several technical measures to promote robustness in treatment-effect prediction	Preprocessing, calibration, and evaluation procedures are used to improve reliability, and calibration " <i>is used to improve predictive reliability</i> ".	These steps help stabilize predictions despite heterogeneous data quality.
DEN-CER 6	The system incorporates basic privacy and security safeguards but lacks a fully developed governance framework.	While encryption and access controls exist, governance responsibilities " <i>are not fully defined or documented</i> ".	The system includes basic protections but lacks the formal governance structures required for full regulatory alignment.
DEN-CER 7	The system supports clinician decision-making but does not include mechanisms to prevent overreliance.	Clinicians remain responsible for decisions, yet the system " <i>does not include structures to counter overreliance</i> ".	Without oversight mechanisms, there is a risk that clinicians may rely too heavily on model outputs.
DEN-CER 8	The system recognizes potential biases but does not	Bias risks are recognized, yet " <i>no evaluation has been conducted on</i>	The absence of fairness assessments limits the system's ability to ensure eq-

	implement formal fairness evaluation or mitigation.	<i>underrepresented groups</i> " and no mitigation procedures exist.	uitable performance across populations.
DEN-CER 9	The system offers limited accountability structures, particularly regarding versioning and error tracking.	Documentation supports reproducibility, but " <i>no structured MLOps process or formal audit trail exists</i> ".	Without versioning, error tracking, or oversight structures, accountability remains limited.

Appendix 2: Interview Questions and Responses

The interview was conducted on February 7, 2026. The interviewee was *Copilot*, a generative AI-based chatbot developed by *Microsoft*. Before the interview, *Copilot* received training on the topics of the thesis using the prompt engineering method, as described by Ye et al. (2024).

Background and Context

Question 1: Could you describe your role and responsibilities in healthcare product innovation and development?

"I work as a project manager overseeing cross-functional development projects for digital health and medical device solutions. My responsibilities include coordinating multidisciplinary teams, managing timelines and regulatory deliverables, ensuring compliance with quality management systems, and facilitating communication between engineering, clinical, and regulatory stakeholders. I also support risk management activities and oversee the integration of new technologies, including AI components, into our product development pipeline."

Question 2: What kinds of AI systems or data-driven tools have you encountered in your work?

"In our projects, we use AI-enabled tools for predictive analytics, resource forecasting, and early risk identification. On the product side, I've been involved in projects where machine-learning models support clinical decision-making or automate parts of diagnostic workflows. We also use data-driven tools for document automation, requirements management, and post-market surveillance, where AI helps detect safety signals from large datasets."

Question 3: How would you describe your familiarity with AI governance or trustworthiness concepts?

“I’m familiar with the general principles of trustworthy AI, such as transparency, explainability, fairness, and human oversight. In healthcare, these principles are closely tied to regulatory expectations, so we consider them when planning validation activities or documenting compliance. I’m aware of frameworks like the EU AI Act and ISO standards related to AI management systems, and I’ve participated in internal discussions about how to operationalize ethical and regulatory requirements in our development processes.”

AI Use in Project Management

Question 1: In your experience, how is AI currently used in project management activities within healthcare product innovation and development?

“AI is increasingly used to support analytical and administrative aspects of project management. For example, predictive analytics tools help forecast project timelines, identify potential bottlenecks, and estimate resource needs based on historical data. We also use AI-enabled systems for document management, such as automated extraction of requirements or classification of regulatory documents. In some cases, AI assists in risk identification by scanning large datasets for patterns that might indicate emerging issues. While AI does not replace project managers, it augments decision-making by providing data-driven insights.”

Question 2: Which project management tasks do you believe are most affected by AI tools?

“Planning and risk management are the areas where AI has the most noticeable impact. AI tools can analyze past project performance to generate more accurate schedules or highlight risks that might not be obvious to human reviewers. Resource allocation is another area where AI is helpful, especially in complex projects with multiple dependencies. Additionally, AI supports documentation-heavy tasks, such as preparing

regulatory submissions or maintaining traceability matrices, which are essential in healthcare development.”

Question 3: What benefits have you observed or expect from using AI in project management?

“The main benefit is improved decision quality. AI can process large volumes of data much faster than humans and identify trends that might otherwise go unnoticed. This leads to more realistic planning and earlier detection of risks. Another benefit is efficiency: automating repetitive tasks frees up project managers to focus on strategic issues and stakeholder communication. In healthcare projects, where documentation and compliance are critical, AI can also reduce human error and improve consistency.”

Question 4: What challenges or limitations have you encountered?

“One challenge is the reliability of AI-generated insights. If the underlying data is incomplete or biased, the recommendations may be misleading. Another limitation is explainability: AI tools sometimes provide outputs without clear justification, which makes it difficult to use them in regulated environments where traceability is required. Integration with existing systems can also be difficult, especially in organizations with legacy infrastructure. Finally, there is a cultural challenge — some team members are hesitant to trust AI tools or fear that automation might reduce their role.”

Perceptions of Trustworthy AI

Question 1: When you think about ‘trustworthy AI,’ what aspects or principles come to mind?

“For me, trustworthy AI means systems that behave reliably, transparently, and in ways that align with ethical and regulatory expectations. Key principles include transparency about how the model works, explainability of its outputs, fairness in how it treats different patient groups, and strong data governance. Human oversight is also essential — especially in healthcare, AI should support decisions, not replace clinical or project judgment.”

Question 2: How important do you consider issues such as transparency, explainability, fairness, or human oversight in AI systems used in healthcare projects?

“These issues are absolutely critical. In healthcare product development, every decision must be traceable and justified, so an AI system that cannot explain its reasoning is difficult to integrate into regulated workflows. Fairness is also important because biased outputs could lead to unsafe or inequitable outcomes. Human oversight is non-negotiable — regulators expect it, and project teams rely on it to maintain accountability.”

Question 3: Have you encountered situations where trustworthiness concerns influenced decisions about AI adoption or use?

“Yes, there have been cases where we decided not to use certain AI tools because the vendor could not provide sufficient transparency about the model or its training data. In other situations, concerns about data privacy or unclear validation results led us to delay adoption until we had stronger evidence. Trustworthiness concerns often slow down or reshape how AI is integrated into our processes.”

Risks and Ethical Considerations

Question 1: What kinds of risks do you associate with using AI in healthcare product innovation and development projects?

“The main risks relate to reliability and unintended consequences. If an AI model is trained on incomplete or biased data, it may produce misleading recommendations that affect project decisions. There is also the risk of over-reliance — teams might trust AI outputs too much without critically evaluating them. From a regulatory perspective, insufficient documentation or validation can jeopardize compliance.”

Q2. How do you think AI might influence decision-making quality or accountability in project teams?

“AI can improve decision quality by providing data-driven insights, but it can also complicate accountability. If a project decision is based on an AI recommendation, it may

be unclear who is responsible if something goes wrong — the project manager, the data scientist, or the tool provider. This ambiguity can create governance challenges. It's important to maintain clear human responsibility for all decisions."

Question 3: Are there ethical or societal concerns you believe should be considered when integrating AI into healthcare project workflows?

"Absolutely. Ethical concerns include patient privacy, potential bias in algorithms, and the risk of reducing human judgment in critical decisions. Societal concerns involve trust — both within the organization and among end-users or patients. If AI is perceived as opaque or unfair, it can undermine confidence in the final product. We also need to consider long-term impacts, such as how AI might shift roles or competencies within project teams."

Question 4: How do regulatory requirements (e.g., MDR, EU AI Act) shape your view of AI risks?

"Regulatory requirements make the risks more concrete. The MDR and the EU AI Act emphasize transparency, documentation, and human oversight, which means any AI system used in development must meet high standards of traceability and robustness. These regulations highlight that AI is not just a technical tool but a regulated component that must be managed with the same rigor as any other safety-critical element. They also make it clear that insufficient governance is itself a risk."

Integration of Trustworthiness Assessment into Project Management

Question 1: Do you think trustworthiness assessment should be part of project management processes? Why or why not?

"Yes, I believe trustworthiness assessment should be embedded into project management, especially in healthcare. AI introduces risks that traditional project management methods do not fully address, such as algorithmic bias, data quality issues, and explainability gaps. If these concerns are not managed proactively, they can jeopardize regulatory compliance and patient safety. Integrating trustworthiness assessment ensures

that ethical, technical, and regulatory considerations are systematically evaluated throughout the project lifecycle.”

Question 2: At which stages of the project lifecycle do you think AI trustworthiness should be evaluated?

“Ideally, trustworthiness should be assessed at multiple stages. During initiation, it helps determine whether the AI use case is appropriate and feasible. In planning, it informs risk management and quality requirements. During execution, it guides validation and monitoring activities. In the monitoring and controlling phase, it ensures that the AI system continues to behave as expected. Finally, at project closure, it supports documentation and post-market planning. So, it’s not a one-time activity — it’s a continuous process.”

Question 3: What practical challenges might arise when integrating trustworthiness assessments into project workflows?

“One challenge is the lack of clear ownership — it’s not always obvious whether responsibility lies with data scientists, project managers, or regulatory teams. Another challenge is the additional workload: trustworthiness assessments require documentation, testing, and stakeholder engagement, which can extend timelines. There may also be gaps in organizational expertise, especially around ethical evaluation or socio-technical analysis. Finally, integrating new assessment methods into existing quality systems can be difficult.”

Question 4: What kinds of tools, frameworks, or guidelines would help project managers evaluate AI trustworthiness effectively?

“Project managers would benefit from structured frameworks that translate high-level ethical principles into actionable steps. Tools like checklists, risk matrices, and standardized documentation templates would be useful. Clear guidelines on how to evaluate explainability, data governance, and human oversight would also help. Ideally, these

tools should align with regulatory expectations, such as the EU AI Act or MDR, so that trustworthiness assessment supports compliance rather than adding redundant work.”

Z-Inspection® Method

Question 1: Are you familiar with the Z-Inspection® method for assessing trustworthy AI?

“I’m familiar with the general idea. Z-Inspection® is a structured process for evaluating the ethical and societal impact of AI systems. It focuses on contextual analysis, stakeholder involvement, and identifying socio-technical risks. I haven’t used it directly, but I understand its purpose and approach.”

Question 2: Based on your understanding, how applicable do you think Z-Inspection® is to healthcare product innovation and development projects?

“I think it’s highly applicable. Healthcare is a domain where ethical and societal considerations are central, and Z-Inspection® provides a systematic way to evaluate them. Its emphasis on context is particularly valuable because AI systems behave differently depending on how and where they are deployed. The method also aligns well with regulatory expectations for transparency and risk management.”

Question 3: What strengths or limitations do you see in using Z-Inspection® within project management?

“A major strength is that it brings structure to ethical evaluation, which is often handled informally. It also encourages multidisciplinary collaboration, which is essential in healthcare. However, one limitation is that it can be resource-intensive. It requires time, expertise, and stakeholder engagement, which may be challenging in fast-paced projects. Another limitation is that it may need adaptation to fit into existing project management frameworks and quality systems.”

Question 4: How might such a method support compliance, risk management, or decision-making?

“Z-Inspection® can strengthen compliance by ensuring that ethical and societal risks are documented and addressed systematically. It supports risk management by identifying issues that might not emerge through technical testing alone. For decision-making, it provides a structured basis for evaluating whether an AI system is appropriate for deployment and under what conditions. This can help project managers justify decisions to regulators, auditors, and internal stakeholders.”

Organizational Readiness and Capability

Question 1: How prepared do you think your organization (or the industry in general) is to assess AI trustworthiness systematically?

“I would say we are partially prepared. There is awareness of the importance of trustworthy AI, and some processes exist for data governance and validation. However, systematic assessment is still emerging. Many organizations lack formal frameworks or dedicated roles for AI governance. The industry is moving in the right direction, but there is still a maturity gap.”

Question 2: What competencies or roles do you think are needed to support trustworthy AI assessments in projects?

“We need a combination of technical, regulatory, and ethical expertise. Data scientists and engineers understand the models, but project managers and regulatory specialists understand compliance requirements. We also need people with expertise in ethics, human-computer interaction, and socio-technical analysis. Ideally, organizations should establish cross-functional AI governance teams that bring these competencies together.”

Question 3: What barriers (cultural, technical, regulatory, resource-related) might hinder adoption?

“Culturally, there may be resistance to change or skepticism about the value of trustworthiness assessments. Technically, organizations may lack the tools or infrastructure to evaluate AI systems thoroughly. Regulatory uncertainty can also be a barrier — teams may not know how to interpret new requirements like the EU AI Act. Resource constraints are another issue: trustworthiness assessments require time, expertise, and budget, which may be limited in smaller organizations.”

Future Outlook

Question 1: How do you see the role of AI evolving in project management for healthcare product innovation and development?

“I expect AI to become increasingly embedded in project management workflows, especially for predictive analytics, risk forecasting, and documentation automation. As datasets grow and tools mature, AI will likely shift from being a support function to becoming a strategic partner in decision-making. However, I don’t see AI replacing project managers — instead, it will augment their capabilities by providing deeper insights and reducing administrative burden. In healthcare, AI will also play a larger role in ensuring compliance by helping teams maintain traceability and consistency across complex documentation requirements.”

Question 2: What changes do you think will be necessary to ensure AI is used responsibly and effectively?

“Organizations will need clearer governance structures and standardized processes for evaluating AI systems. This includes integrating trustworthiness assessments into quality management systems, establishing cross-functional AI oversight committees, and ensuring that teams have the right competencies to understand and challenge AI outputs. Training will be essential — not just for technical staff, but for project managers and regulatory professionals as well. Additionally, AI tools will need to become more transparent and interoperable so they can fit into existing workflows without creating new risks.”

Question 3: If you could recommend one improvement to current practices regarding AI governance in projects, what would it be?

“I would recommend formalizing AI governance as part of the project lifecycle rather than treating it as an optional or ad-hoc activity. This means defining clear checkpoints for evaluating trustworthiness, documenting decisions, and involving the right stakeholders at the right time. A structured approach would help reduce uncertainty, improve accountability, and ensure that ethical and regulatory considerations are addressed proactively rather than reactively.”

Question 4: Is there anything you believe organizations should start preparing for now, given the direction AI regulation and technology are heading?

“Organizations should start preparing for stricter regulatory expectations, especially with the EU AI Act coming into force. This will require better documentation practices, more robust validation procedures, and clearer accountability structures. They should also invest in data quality and governance, since these will become foundational to both compliance and performance. Finally, organizations should prepare for cultural change — as AI becomes more prevalent, teams will need to adapt their ways of working, collaborate more closely across disciplines, and develop a shared understanding of what trustworthy AI means in practice.”

Appendix 3: Trustworthy AI Assessment Template for Project Managers

This **Appendix** includes a project management-ready template that can be plugged directly into project's workflows. The template is clear, simple, and structured so that it fits naturally into common project management materials (charters, stage-gates, risk logs, sprint ceremonies, etc.) while still adhering to the main HLEG trustworthy AI requirements. For more information, always check the ethics guidelines (AI HLEG, 2020) or their assessment list (ALTAI, 2020).

The template can be treated as a reusable template for any AI-driven innovation and development project. The template was developed in this thesis.

Project Information

Project Name:

Project Owner or Sponsor:

AI System Description:

Intended Use and Users:

Lifecycle Stage: (Initiation / Planning / Execution / Monitoring / Closure)

Assessment Date:

Assessor(s):

1. Human Agency and Oversight

Objectives

Ensure the AI system supports human decision-making and includes appropriate oversight mechanisms.

Checklist

- Users are informed when interacting with AI
- System capabilities and limitations are clearly documented
- Human-in-the-loop / on-the-loop / in-command approach defined
- Mechanisms exist to override or halt the system

User training or guidance materials prepared

Risks & Mitigations

Identified Risks:

Mitigation Actions:

Owner:

Due Date:

2. Technical Robustness and Safety

Objectives

Ensure the system is trustworthy, secure, and resilient to failures.

Checklist

Accuracy, trustworthy, and reproducibility tested

Stress tests and adversarial tests performed

Fallback and fail-safe procedures defined

Error handling and recovery processes documented

Monitoring plan for model drift and anomalies

Risks & Mitigations

Identified Risks:

Mitigation Actions:

Owner:

Due Date:

3. Privacy and Data Governance

Objectives

Ensuring data is handled trustworthy, lawfully, and with strong governance.

Checklist

- Data sources documented and validated
- Data quality and integrity checks performed
- Privacy impact assessment completed
- Access controls and permissions defined
- Data retention and deletion policies in place

Risks & Mitigations

Identified Risks:

Mitigation Actions:

Owner:

Due Date:

4. Transparency**Objectives**

Ensure the system is understandable, traceable, and explainable to relevant stakeholders.

Checklist

- Model documentation (data, design, logic) completed
- Traceability mechanisms implemented
- Explainability approach defined for each stakeholder group
- User-facing explanations available
- Communication plan for system purpose and limitations

Risks & Mitigations

Identified Risks:

Mitigation Actions:

Owner:

Due Date:

5. Diversity, Non-Discrimination and Fairness

Objectives

Ensure the system treats individuals and groups fairly and inclusively.

Checklist

- Bias assessment performed on data and model outputs
- Mitigation strategies for identified biases implemented
- Accessibility requirements addressed
- Stakeholder groups consulted
- Inclusive design principles applied

Risks & Mitigations

Identified Risks:

Mitigation Actions:

Owner:

Due Date:

6. Societal and Environmental Well-Being

Objectives

Ensure the system contributes positively to society and minimizes environmental impact.

Checklist

- Societal impact assessment completed
- Environmental impact (e.g., energy use) evaluated
- Long-term implications considered
- Potential harm to ecosystems or communities assessed
- Alignment with organizational values verified

Risks & Mitigations

Identified Risks:

Mitigation Actions:

Owner:

Due Date:

7. Accountability

Objectives

Ensure clear responsibility, auditability, and redress mechanisms.

Checklist

- Roles and responsibilities defined
- Audit processes established
- Documentation supports internal/external review
- Redress mechanisms available for affected users
- Compliance with legal and ethical standards verified

Risks & Mitigations

Identified Risks:

Mitigation Actions:

Owner:

Due Date:

Summary and Go or No-Go Recommendation

Overall Assessment: (Compliant / Partially Compliant / Non-Compliant)

Key Issues Blocking Progress:

Decision: (Proceed / Proceed with Conditions / Hold / Stop)

Decision Maker:

Date: