A generic & customizable methodology based on quality gates towards certifiable AI in medicine

Miriam Katrin Sophie Elia

Angaben zur Veröffentlichung / Publication details:

Elia, Miriam Katrin Sophie. 2025. “A generic & customizable methodology based on quality gates towards certifiable AI in medicine.” Augsburg: Universität Augsburg.

A generic & customizable Methodology based on Quality Gates towards Certifiable AI in Medicine

Miriam Katrin Sophie Elia

Dissertation
for the degree of
Doctor of Natural Sciences (Dr. rer. nat.)

2025

University of Augsburg
Department of Computer Science
Software Methodologies for Distributed Systems

A generic and customizable Methodology based on Quality Gates towards Certifiable AI in Medicine

Supervisor:	**Prof. Dr. Bernhard L. Bauer**, Software Methodologies for Distributed Systems, University of Augsburg, Germany
Advisor:	**Prof. Dr. Frank Kramer**, IT Infrastructures for Translational Medical Research, University of Augsburg, Germany
Advisor:	**Prof. Dr. Alexander Schiendorfer**, AI-based Optimization in Automotive Production, Ingolstadt University of Applied Sciences, Germany

Day of Defense: October 29th 2025

Public AI tools were used to enhance this text.

The whole is greater than the sum of its parts

*A*ristotle

Abstract

Artificial Intelligence (AI) technologies are entering the real world at an accelerated pace, offering remarkable benefits across domains. However, especially in high-risk areas such as healthcare, their deployment also presents significant ethical, legal, and societal risks. The European Union's AI Act (EU AI Act) provides regulatory oversight in such contexts and aims to ensure fundamental rights, safety, and health [Fut24]. While the legislation offers valuable guidance, moving towards Responsible AI (RAI) [DRN23] requires more than the formulation of principles. It demands operational structures that support quality assessment, transparency, and traceability across the entire AI lifecycle. In this context, AI Quality Management (AI QM) counts as a key enabler for aligning AI system creation with responsible innovation and legal compliance. Despite the increasing availability of guidelines, best practices, and standards, stakeholders face persistent challenges in applying these partly abstract principles to concrete system design decisions. This is particularly difficult due to the intricate nature of AI systems: they are opaque, data-driven, stochastic, and use case-specific, involving interdependent design choices across multiple lifecycle stages. Further, relevant design information is typically scattered across disciplines and formats, limiting reusability and collective learning.

This thesis addresses these challenges by proposing a structured and extensible blueprint to organizing AI design knowledge for RAI system creation, resulting in the *Generic and Customizable Methodology based on Quality Gates towards Certifiable AI in Medicine* (MQG4AI). We focus on the medical domain through our use cases, but the blueprint is intended to be generalizable across domains. MQG4AI introduces a dynamic Information Management (IM) structure that supports collective and transparent AI design by allowing for the organization and contextualization of new and existing contributions from research and industry. The blueprint integrates principles from Design Science Research (DSR) [vJ20], aiming to continuously bridge the gap between high-level RAI guidance and context-specific AI system implementations. At the core of MQG4AI lies the concept of Quality Gates (QG), adapted from traditional software engineering [Fil06] [Flo08] [GM09] [HS09], aiming to support quality-assured decisions throughout the AI lifecycle. QGs guide lifecycle design in a hierarchical manner, starting from generic lifecycle processes (e.g. the model development stage) to use-case-specific configurations (e.g. performance metrics in medical image segmentation). Additionally, they enable the inclusion of interdependencies between design decisions, as well as linking contextual information with the AI lifecycle. The blueprint incorporates ethical considerations and operational requirements from the EU AI Act, particularly Article 17 (Quality Management) and Article 9 (Risk Management) [Fut24]. Further, we exemplify aligning AI risks with the Trustworthy AI (TAI) principles outlined by the European Commission [Eur19] [Hig20].

Within our contribution, special emphasis is placed on the development phase of AI models, identifying which information should be extracted or integrated to support downstream (e.g. deployment) and upstream (e.g. data acquisition) activities. Moreover, MQG4AI functions as a knowledge-sharing tool, facilitating the on-going aggregation and practical accessibility of RAI design knowledge in a decentralized manner. Following DSR, the MQG4AI blueprint allows for the continuous communication between RAI lifecycle design knowledge (MQG4DK) and individual AI projects (MQG4A). These templates, which are derived from the global blueprint, are envisioned to provide use case-adapted guidelines for high-quality AI lifecycle design, promoting traceability from a conceptual perspective. Thus, MQG4AI is envisioned to enable shared, flexible, and use case-oriented lifecycle design for actively involved stakeholders. It supports the reuse of lessons learned, thereby contributing to the long-term goal of enhancing AI literacy, as emphasized in Article 4 of the EU AI Act [Fut24]. A visualization of the MQG4AI blueprint is available on GitHub.[1]

In conclusion, this thesis offers a conceptual and practical foundation for enabling RAI through structured IM along the AI lifecycle. By supporting continuous, collaborative, and compliant AI development practices, MQG4AI contributes to the realization of trustworthy, high-quality AI systems capable of meeting both, ethical expectations and regulatory demands. This work aspires to empower development teams and regulatory actors alike, closing the gap between AI's technical complexity and the societal values it must uphold.

[1] https://github.com/miriamelia/MQG4AI/blob/main/README.md

Zusammenfassung

Künstliche Intelligenz (KI) hat großes Potenzial für vielfältige Anwendungsbereiche und hält mit zunehmender Geschwindigkeit Einzug in reale Anwendungskontexte. Besonders in Hochrisikodomänen wie dem Gesundheitswesen birgt der Einsatz dieser Technologie jedoch ethische, rechtliche und gesellschaftliche Risiken. Die europäische KI-Verordnung, der EU AI Act, formuliert Anforderungen für die regulatorische Aufsicht in solchen Kontexten und verfolgt das Ziel, Grundrechte, Sicherheit und Gesundheit zu schützen [Fut24]. Obgleich dort wertvolle Leitlinien beschrieben werden, geht der Weg zu verantwortungsvoller KI (RAI) [DRN23] in der realen Welt über die bloße Formulierung von Anforderungen hinaus. Es bedarf operativer Strukturen, die Qualitätsbewertung, Transparenz und Nachvollziehbarkeit über den gesamten KI-Lebenszyklus hinweg unterstützen. In diesem Zusammenhang eröffnet Qualitätsmanagement (QM) für KI zentrale Möglichkeiten, die Entwicklung von KI-Systemen im Einklang mit verantwortungsvoller Innovation und rechtlicher Konformität zu gestalten. Trotz der zunehmenden Verfügbarkeit von Leitlinien, Best Practices und Standards stehen die Akteur:innen vor anhaltenden Herausforderungen, diese teils abstrakten Regelungen auf konkrete Designentscheidungen anzuwenden. Dies wird durch die komplexe Natur von KI-Systemen zusätzlich erschwert: Sie sind intransparent, datengetrieben, stochastisch und anwendungsspezifisch. Hinzu kommt dass Designentscheidungen miteinander über verschiedene Phasen des Lebenszyklus hinweg verflochten sind und somit Abhängigkeiten aufweisen. Auch sind relevante Designinformationen oft disziplinübergreifend und in unterschiedlichen Formaten verstreut, was Wiederverwendbarkeit und kollektives Lernen einschränkt.

Ausgehend von diesen Herausforderungen widmet sich die vorliegende Arbeit einem strukturierten und erweiterbaren Ansatz zur Organisation von KI-Designwissen, um die Entwicklung verantwortungsvoller KI-Systeme gezielt zu unterstützen. Dabei stellen wir eine generische und anpassbare Methodik auf Basis von *Quality Gates* vor, mit dem Ziel zertifizierbare KI im medizinischen Kontext zu unterstützen (MQG4AI). Zwar konzentrieren sich unsere Anwendungsfälle auf das Gesundheitswesen, doch der *Blueprint* ist so konzipiert, dass er auf verschiedene Sektoren übertragbar sein soll. MQG4AI unterstützt die strukturierte Organisation und kontextbezogene Einordnung neuer, sowie bestehender Design-Beiträge aus Forschung und Industrie entlang des KI-Lebenszyklus durch eine integrierte Informationsmanagement-Struktur (IM). Dabei steht die Förderung eines kollaborativen und transparenten KI-Designs im Vordergrund. Unser Vorschlag basiert auf Prinzipien des *Design Science Research* (DSR) [vJ20] und verfolgt das Ziel, die Lücke zwischen abstrakten Leitlinien und kontextspezifischen KI-Implementierungen systematisch zu schließen. Im Zentrum von MQG4AI stehen *Quality Gates* (QG), ein Konzept, welches aus der klassischen

Softwareentwicklung adaptiert wurde [Fil06] [Flo08] [GM09] [HS09], mit dem Ziel, qualitätsgesicherte Entscheidungen entlang des KI-Lebenszyklus zu unterstützen. QGs leiten die Gestaltung des Lebenszyklus hierarchisch an, von generischen Lebenszyklusprozessen (z.B. der Modellentwicklungsphase) bis hin zu anwendungsspezifischen Konfigurationen (z.B. Performanzmetriken in der medizinischen Bildsegmentierung). Darüber hinaus ermöglichen sie die Einbeziehung von Wechselwirkungen zwischen Designentscheidungen, sowie die Verknüpfung kontextueller Informationen mit dem KI-Lebenszyklus. Insgesamt integriert MQG4AI ethische Anforderungen und operationale Vorgaben aus dem EU AI Act, insbesondere Artikel 17 (Qualitätsmanagement) und Artikel 9 (Risikomanagement) [Fut24]. Zudem wird exemplarisch dargestellt, wie KI-Risiken anhand der Prinzipien für vertrauenswürdige KI (TAI) der Europäischen Kommission strukturiert werden können [Eur19] [Hig20]. Ein besonderer Fokus liegt auf der Entwicklungsphase von KI-Modellen, wobei illustrativ aufgezeigt wird, welche Informationen idealerweise extrahiert oder integriert werden sollten, um nachgelagerte (z. B. Deployment) und vorgelagerte (z. B. Datenerhebung) Aktivitäten zu unterstützen. Darüber hinaus unterstützt MQG4AI den Wissensaustausch und ermöglicht eine kontinuierliche, praxisnahe Zugänglichkeit von RAI-Designwissen in dezentraler Form. Im Sinne des DSR-Ansatzes, erlaubt dieser Entwurf eine fortlaufende Wechselwirkung zwischen Designwissen für eine verantwortungsvolle Gestaltung des KI-Lebenszyklus (MQG4DK) und individuellen KI-Projekten (MQG4A). Die dadurch abgeleiteten Designvorlagen sollen kontextspezifische Leitlinien für eine qualitativ hochwertige Gestaltung des KI-Lebenszyklus bereitstellen und dessen Nachvollziehbarkeit aus konzeptioneller Sicht fördern. MQG4AI zielt somit darauf ab, eine kollektive, flexible und anwendungsbezogene Lebenszyklusgestaltung für aktiv beteiligte Akteur:innen zu ermöglichen. Dabei wird die Wiederverwendung gewonnener Erkenntnisse unterstützt, was langfristig zur Stärkung der KI-Kompetenz beiträgt, wie sie in Artikel 4 des EU AI Acts betont wird [Fut24]. Eine Visualisierung von MQG4AI ist auf GitHub verfügbar.[2]

Abschließend liefert diese Arbeit eine konzeptionelle und praxisorientierte Grundlage, um verantwortungsvolle KI durch ein strukturiertes IM entlang des KI-Lebenszyklus zu ermöglichen. Durch die Unterstützung kontinuierlicher, kollaborativer und konformer KI-Entwicklungspraktiken trägt MQG4AI zur Realisierung vertrauenswürdiger und qualitativ hochwertiger KI-Systeme bei, die sowohl ethischen Erwartungen als auch regulatorischen Anforderungen gerecht werden. Diese Arbeit hat zum Ziel, sowohl Entwicklungsteams als auch Regulierungsakteur:innen zu unterstützen und die Lücke zwischen der technischen Komplexität von KI und den gesellschaftlichen Werten, die sie im realen Einsatz bewahren soll, zu überbrücken.

[2] `https://github.com/miriamelia/MQG4AI/blob/main/README.md`

Acknowledgments

The journey of creating this thesis has profoundly shaped who I am today, and I could not have done it alone, nor would I have wanted to. For all the guidance, encouragement, and heartfelt companionship, I extend my deepest gratitude.

I am especially grateful to my supervisor, Prof. Dr. Bernhard Bauer, for his guidance, deep insights, and flexibility. I sincerely thank my reviewers, Prof. Dr. Frank Kramer and Prof. Dr. Alexander Schiendorfer, as well as my examiners Prof. Dr. Elisabeth André and Prof. Dr. Annemarie Friedrich for taking the time to engage critically with this work. Many thanks to my colleagues, fellow PhD students, and Post-docs at the University of Augsburg for our shared time. Special thanks to Dr. Dominik Müller, who motivated me from the very beginning.

To my mentors, Prof. Dr. Sophie Weerts, Dr. Jauwairia Nasir, and Emma Grönvik Möller, thank you so much for your insight and encouragement throughout this process. A warm thank-you to Prof. Dr. med. Nina Ditsch and Anna Ohnmeiß, to my peer group, Natalie Rohrmoser, Victoria Fincke, Anna Weininger, and all other contributors, for enabling this valuable mentorship experience and growing together. To all women scientists, grateful for having met so many of you. To Dr. Ronald A. Crutcher, thank you for opening my eyes to pursue mentoring.

To everyone in science communication and STEM education, to Dr. Marietta Menner, Julia Thurner-Irmler and Fabio Hellmann, thank you for your positive energy and dedication. Thinking of the people we reach continues to inspire me. Heartfelt thanks to our *Achalasia Team*, Dr. med. Sandra Nagl, Dr. med. Alanna Ebigbo, Tamara Krafft, and all the motivated Students, for creating a meaningful tool and unforgettable memories. Thank you, Dr. Susanne Wiedemann and Prof. Dr. Kirsten Ostherr for sharing your insight and a wonderful first podiums discussion. To Prof. Dr. Rafael Mayoral, thank you for my first guest lecture.

To the global scientific community, your work drives me. Special thanks to the *Center for Responsible AI Technologies*, and the *LIFEDATA Team* for fostering interdisciplinary exchange, to Prof. Dr. Kerstin Schlögl-Flierl, Paula Ziethmann, Prof. Dr. Sarah Friedrich-Welz, Alina Lorenz, Dr. Julia Krumme, Dr. Fabian Ripke, and Dr. Marius Nann for sharing your valuable knowledge. My deepest gratitude to Prof. Dr. Esteban García-Cuesta, Alba Lopez and Katherin Corredor for your support, expertise, and our inspiring dialogues, profoundly shaping this work.

Last but not least, to my parents and family, thank you for the gift of this life and for your unwavering support through everything. I love you. To my friends and soul sisters, thank you for your laughter, presence, and encouragement. And to nature, for being my quiet space to think, walk, and let new ideas grow...

Contents

III. APPLICATION AND EVALUATION 236

Part I.

INTRODUCTION AND FOUNDATIONS

1 Introduction

Imagine a world, where healthcare is distributed equally across the globe, while providing the highest possible quality to everyone in need. Or, students receive education tailored to their individual needs, and the industry no longer requires humans to execute repetitive tasks. The list is endless. This vision is not necessarily an abstract dream, but can become reality thanks to a combination of technological advancements transforming the world as we know it, highlighting the crucial role of Artificial Intelligence (AI). However, implementing this vision is not trivial and currently, a multitude of complex challenges prevails. AI's novelty to the extent that we experience it today, highlighting Deep Neural Networks (DNN) and the technique's inherent dynamics such as its non-deterministic and opaque character, complicate a trustworthy real-world integration of intelligent systems. We need a responsible fusion of AI with societies and across sectors, otherwise we might wake up in a world one day, where technology dictates our choices, already existing biases are amplified, and trust in critical systems erodes. It is our responsibility to promote a beneficial outcome for everyone, and currently, we are in the beginning stages of defining established procedures towards realizing Responsible AI (RAI) for the multitude of possible use cases across domains.

With this thesis, we aim to contribute a holistic approach for actively involved stakeholders that continuously ensures the desired quality of AI system design through formatted Information Management (IM) along the AI lifecycle. Long-term, the introduced building blocks are envisioned to be implemented as a tool. Part I motivates our proposed methodology, as well as outlines its foundations, that form the basis for the proposed conceptual setup in part II, which is tested on concrete use cases in part III.

1.1. Motivation & Vision

The beneficial character of AI is best illustrated with a real-world example. For instance, focusing on the medical domain, where our use cases are situated, *AlphaFold* [SA20] [PJ21b] has achieved astonishing results in predicting protein structures based on amino acid sequences, which is a complex task due to millions of possible permutations that have not all been discovered as of today. Proteins are building blocks of life and crucial for the development of medication and vaccines, among other things. Consequently, deepening knowledge on their setup enhances overall healthcare, resulting in more personalized care options – a process which, when limited to experimental outputs, requires years. Another example where AI demonstrates significant potential is in radiology and medical image segmentation, use cases that promise to alleviate the strain on overburdened healthcare systems once deployed. However, this technology is not without risks and needs to be implemented, tested and monitored in a reliable manner towards outcomes that benefit humanity and the planet. Therefore, it is important to remember that the true success of any intelligent application can only be measured in the real-world.

This section motivates our proposed Generic and Customizable Methodology based on Quality Gates towards Certifiable AI in Medicine (MQG4AI), its context, vision, as well as concretizes objectives of our contribution. Finally, this thesis is outlined, as well as our scientific participation, including related publications and supervised thesis.

1.1.1. The EU AI Act

On August 1st 2024, the legislation around this fascinating technology within the European Union (EU) came into force, i.e. the AI Act. It aims to ensure product safety of intelligent systems entering the European market while safeguarding fundamental rights, health, and safety of individuals in the European market. AI Systems comprise any application that embeds one or more AI models. More information on the legislative and ethical landscape, as well as considerations on how to implement it can be found in section 3.1. The AI Act classifies intelligent systems into four levels of risk: minimal, limited-, high-, and unacceptable risk with augmenting requirements until prohibition. The present thesis aims to contribute to the required Quality Management System (QMS), as outlined in Article 17 for high-risk AI, focusing on how to implement Trustworthy AI (TAI) for DNNs in accordance with the EU AI Act. [Fut24]

Trustworthy AI In Recital 27, the Act states: "While the risk-based approach is the basis for a proportionate and effective set of binding rules, it is important to recall the 2019 Ethics guidelines for trustworthy AI [...]" [Fut24]. The Act's ethical foundation, as introduced by a High-Level Expert Group appointed by the European Commission (HLEG) [Eur19], outlines "[...] seven non-binding ethical principles for AI which are intended to help ensure that AI is trustworthy and ethically sound" [Fut24]: Human Agency and Oversight (the application of

AI should respect human autonomy); Technical Robustness and Safety (AI systems must be dependable and resilient to changes so that harm is minimized); Privacy and Data Governance (the right to privacy needs to be respected and qualitative data used); Transparency (AI system components need to be traceable, explained as much as possible and communicated to humans); Diversity, Non-discrimination and Fairness (AI systems need to be designed in a way that avoids unfair biases and respects human diversity); Societal and Environmental Well-being (consequences of AI usage and development should respect society and the planet); and Accountability (regarding the realization of all TAI criteria, responsibilities need to be implemented so that intelligent systems are auditable and accountable [DRN23, 18]. We aim to contribute through this angle with this thesis). [Eur19]

AI QM Process The implementation of compliant systems, including these rather abstract TAI qualities, is carried out by the Provider through a comprehensive AI QMS, as described in Article 17. It comprises 13 requirements, including a Risk Management System (RMS), as defined in Article 9 [Fut24]. Adding to the complexity of realizing quality assessment for high-risk domains such as medicine, regulatory procedures accompanied by Notified Body are required before to-market-release, including post-market monitoring check-ups. This dependency requires close teamwork early on towards compliance-checks of AI systems. We highlight the pivotal role of provider (domain and tech experts included) and notifying authorities, as defined in Article 3 [Fut24] at the interface of implementation and regulation, when ensuring the required quality of intelligent systems regarding their effects on the user and the intended real-world setting.

1.1.2. Towards *Responsible AI*

The interface between AI ethics and law towards an implementation of Trustworthy AI (TAI) that enables quality assessment is explained very well by Natalia Díaz-Rodríguez et al. [DRN23], and we adopt their transition from TAI to Responsible AI (RAI), which is ethical, lawful, and accountable [DRN23]. The latter focuses on realizing responsibility for the quality of the implementation of individual AI systems, which aims for TAI as the foundation of their behavior in the real-world, where the true success of an intelligent system is measured.

AI Lifecycle Implementing and evaluating the required quality of RAI systems bundles together along the AI lifecycle, i.e. all design decisions that comprise the transition from conceptualization (and research) to production, which includes the humans, or stakeholders that actively participate and assist in the system's creation, namely Contributing Stakeholders.

Lifecycle Planning & AI Information Management Overall, we envision to contribute towards achieving, and maintaining intelligent real-world applications that promote RAI through decentralized AI lifecycle IM and planning. Aiming to assist AI QM, we focus on linking AI risks with lifecycle processes and design decisions towards compliance by design. Concretely, we intend to provide a customizable and generic approach, that enables contributing stakeholders to follow instructions for the domain- and use case-adapted implementation, as well as assessment of RAI projects in the context of certifiable AI in medicine. Long-term, we hope to support a better understanding of this fascinating technology. We are convinced that RAI IM, which incorporates evolving lifecycle design knowledge, comprises a solid foundation for various analysis of the identified methods' interplay. This equally promotes general AI Literacy, as emphasized in Article 4 of the AI Act [Fut24].

1.2. Problem Statement & Research Question

Thanks to AI's inherent dynamics, our incomplete RAI Design Knowledge and the extent of innovation we face today, multiple challenges arise that are required to be addressed towards a RAI real-world integration.

1.2.1. Deep Neural Networks, AI Pitfalls & Use Case-Specificity

We focus on Deep Neural Networks (DNN), a subgroup of Machine Learning (ML) methods that enable "human-like" performance of AI, meaning they can handle complex data and tasks, as well as translate between different dimensionalities, e.g. from text-to-image and vice versa. However, the augmenting complexity of tasks an intelligent system can solve impacts their behavior in multiple ways, which is application-specific and requires a profound AI literacy among contributing stakeholders.

Intricate Deep Neural Networks A major challenge in implementing RAI, comprises shedding light on DNN's output generation, which is incomprehensible to humans. A reliable evaluation and interpretation of model outcomes can only happen indirectly through a combination of metrics and Explainable AI (XAI) techniques, among other methods, which are strongly influenced by the individual use case. Challengingly, DNNs are characterized by opacity through complex interdependencies of billions of parameters, as well as stochastic, and evolutionary behavior with respect to unseen data they encounter. This complexity results in iterative processes for testing towards reliable design decision-making. Intelligent systems that, if built on incomplete knowledge, can result in AI Pitfalls, or methodological and conceptual errors, as introduced in chapter 2. This impedes the realization of RAI.

Use Case-Specificity As a result of this intricate entanglement, every (medical) AI project is based on a unique combination of reasonable design choices for data and model development. Deployment, maintenance, and decommissioning of the intelligent system contribute even more complexity, adding prerequisites for a seamless fusion with (or removal of) the intended real-world setting. This requires a comprehensive and continuous, as well as use case-specific AI lifecycle conceptualization phase. Among others, factors such as data type(s), domain embedded tuning objective(s) and data set distribution(s) impact the respective intelligent system's underlying combination of design decisions, which results in multifaceted possible implementation approaches. Aiming for robust realizations, desirable qualities such as adaptability and generalizability are difficult to implement and assess tailored to individual use cases, and closing the gap to generic guidelines is a non-trivial, but necessary task. Thanks to AI's novelty to the extent that we experience today, we are not aware of all reliable combinations for design decisions. The technology's intricacy results in a multitude of interdependent implementations for possible methods along the AI lifecycle, which requires a profound AI literacy among contributing stakeholders at both sides of compliance assessment.

1.2.2. Standardized Approaches & AI Quality Management

As of now, we are still in the initial phases of analyzing and organizing the immense complexity of possible AI design choices for the multitude of use cases. Especially, for later phases of the AI lifecycle, we currently lack the experience to derive reliable design choices, since especially for high-risk domains, most models are currently under development and have not yet been released to production. Simultaneously, a lot of research is being conducted by the scientific community and the industry alike, and we are advancing at high speed towards ubiquitous AI adoption, which includes covering the implementation of regulatory requirements.

AI Standards A lot of AI knowledge has been, is being and is going to be published in various formats that are scattered around the world/internet, reaching from best practices to standards, which play a crucial role in realizing legislation. The definition of standardized procedures towards reliable design decision-making that are generalizable while providing sufficient space to address use case-specificity requires substantial effort and needs to be aligned with DNNs' inherent dynamics.

Incomplete Coverage The EU AI Act defines 13 requirements for AI QMS in Article 17 [Fut24], which are introduced in more detail in sections 3.1.3.3 and 3.4.1.2. Their implementation for the multitude of existing use cases is not trivial, and no complete coverage of the AI Act through standards currently exists [Eura]. This leaves the concrete realization of QM for individual AI projects an open question, and numerous individuals are engaging with this topic worldwide.

1.2.3. Ethical Questions & Impact of AI

Further, resulting from the disruptive power of AI, and its (unknown) impact on society, a multitude of ethical questions emerge, which are closely related to AI risks. It is to be expected that considerations such as how to address the multifaceted concept of bias, or trade-offs between transparency versus accuracy, will be constant companions of individual high-risk AI projects. This results from contextual information tailored to specific use cases that affects the implementation of the AI lifecycle, including methods to realize TAI and mitigate risks, as well as the consideration of the human influence.

AI Risks & Trustworthiness While general information on AI trustworthiness and risks is being collected [Eur19] [Tab23], we are currently not aware of all possible ways to address AI risks for individual use cases, which can be traced back to a lack of experience with intelligent systems from an application viewpoint. Analogously to the previously introduced open technical challenges, this results from AI's novelty to the extent that we experience today. The central role of RM is outlined in more detail in sections 3.2, 3.3.2, and 3.3.5. Risk analysis, evaluation, and the implementation of control mechanisms, aiming to support an intelligent system that respects safety, fundamental rights, and health [Fut24] along the complete lifecycle, are highly use case-specific. This openness regarding the individual stakeholder's influence can be challenging for compliance, and poses risks in itself if not addressed as part of a continuous AI lifecycle conceptualization. While (semi-)automation offers means to standardize quality, it also introduces risks that must be carefully managed. Striking the right balance is essential, aiming for reliable design decisions and the meaningful involvement of all contributing stakeholders.

Human Decision-Making Human decision-making fundamentally shapes the intelligent system, which includes the user, and every human being has a unique and biased world-view that is formed by factors such as culture or gender. As of now, teams unfortunately tend to lack multidisciplinarity and a diverse background, which could lower biased behavior of the intelligent system through creating awareness from the start of AI lifecycle planning, as an example. In general, the implementation of TAI [Hig20] requires careful consideration when making design choices. A lack of awareness could result in systems that are not developed or assessed in a responsible manner due to a limited understanding of long-term consequences. These dynamics are amplified through an absence of ethics training of contributing stakeholders, and "[...] the difficulty in moving from principles to practice presents a significant challenge to the implementation of ethical guidelines" [PD20, 1].

1.2.4. Research Question

In summary, concentrating on the provider perspective and addressing the interface with compliance for high-risk medical AI, our present contribution explores the overall question:

> *How can we embed ethics and quality by design into structured AI lifecycle information flows that guide the transition from initial ideas, concepts, and data to fully deployed and managed AI systems, in a way that the modeled design knowledge continuously captures technical and conceptual dependencies, as well as socio-ethical considerations to support responsible and compliant implementation for all stakeholders involved?*

1.3. Objectives

This section explores how we envision to address the challenge of designing a holistic approach towards implementing RAI for the multitude of possible use cases, with the long-term vision to establish findings as a tool. Namely, we aim to outline important criteria for a methodology that combines the interface of implementation and regulation to support AI QMS, as defined by the AI Act in Article 17 [Fut24] in a continuous manner, incorporating AI's evolutionary character. Our contribution equally encompasses outlining the extraction of a design workflow, its application to the medical domain and evaluation through concrete use cases.

1.3.1. RAI Methodology – towards Quality by Design

A holistic methodology towards RAI implementation that addresses AI's inherent dynamics is required to integrate ethics and contextual AI system information with continuous AI lifecycle design that is based on responsible design decision-making. Our proposed approach follows the assumption, that an AI system is interpreted as the sum of all underlying lifecycle processes, and design decisions that continuously contribute to the respective AI system's states. They result in the AI's interactions with the intended real-world setting, where the true success, i.e. the system's quality is measured – a non-trivial and highly use case-specific task. All required additional QM information, as introduced in sections 3.1.3.3 and 3.4.1.2, such as resource, and accountability management, as well as relevant documentation can be derived from information on the project-specific lifecycle implementation.

1.3.1.1. Reliable Design Decisions along the AI Lifecycle

Starting from the developer's perspective, and envisioning regulatory compliance by design, which is assessed on process-level [RA23] along the AI lifecycle, we highlight reliable design decision-making. Robust design choices are not trivial thanks to multiple challenges, as previously introduced in section 1.2. Overall, a trustworthy design is only achieved through comprehensive testing and optimization of different methods that are tailored to the intended use, resulting in iterative design decision-making, under consideration of interdependencies. These processes are monitored and documented through continuous AI lifecycle planning, which is supported by comprehensive AI literacy among contributing stakeholders.

AI Lifecycle Planning DNN's inherent dynamics result in the need to identify, correctly implement, and monitor suitable methods along the complete lifecycle tailored to the specific application scenario. Comprehensive AI lifecycle planning that links ethical information with regulatory requirements supports implementing RAI through visibly connecting design choices with supplementary contextual information. A visualization of such dependencies includes linking identified risks/risk controls with implemented design choices, or monitoring the amount of required resources regarding consumed computing power during model training towards a more sustainable usage through optimized processes. Overall, the system's transition from conceptualization to application in the real-world, requires a holistic methodology to allow space for reliable development, continuous monitoring strategies, as well as a compliant data lifecycle as part of comprehensive AI system lifecycle planning, design and testing. Further, highlighting AI's evolutionary character and inherent complexity, as well as our current state of rudimentary AI knowledge, a continuous conceptualization summarizes necessary information for quality assessment. Consequently, a methodology towards RAI should offer a comprehensive lifecycle view that considers all stages in a continuous manner – from the start of conceptualization until the system's decommissioning.

Iterations & Interdependencies A methodology towards RAI needs to address continuous planning and monitoring of multiple lifecycle iterations, and the lifecycle's stages of evolution need to be documented for transparency, reproducibility and traceability. Focusing on model development, for instance, this includes continuous monitoring of design choices surrounding the optimization of chosen hyperparameters. This process includes the comparison of different model development settings compared against a predefined base-line or vanilla approach that are evaluated through a suitable combination of performance metrics, which require to be previously identified. As a result, different stages of the AI lifecycle need to be executed in a repetitive manner until design decisions are finalized. Further, information such as changes in the underlying data distribution, or online learning during application in the real-world can result in necessary updates due to existing interdependencies. Their identification and monitoring through e.g. reasonable thresholds is crucial to implement RAI. All lifecycle stages are interrelated and executed in an iterative manner, which needs to be considered when designing project-specific AI lifecycles. In time, with advancements of standardized methods, lifecycle iterations will be optimized, but in light of AI's complexity and use case-specificity, a certain degree of testing and optimization will always be required. We highlight the importance of continuous AI lifecycle planning and design towards implementing RAI, as starting point for methodology creation.

AI Literacy Finally, addressing AI pitfalls by design to prevent faulty outputs and misuse, necessitates a profound AI literacy of contributing stakeholders, which requires integrated training and Knowledge Management (KM) processes. AI knowledge is relevant to the compliance side, as well, and notified authorities who assess intelligent systems equally need to possess and decode use case-adapted information on DNN's inherent dynamics and how to avoid AI pitfalls. Consequently, a methodology towards RAI should enable AI design KM in alignment with a dynamic real world. It should be understood as a *living blueprint* that aims to continuously capture the state of the art.

1.3.1.2. Dynamic & Qualitative Information Management

AI lifecycle design knowledge on trustworthy, and compliant approaches for all possible use cases will evolve as the technology evolves, and currently, we are at the beginning of AI's on-boarding in the world. This necessitates continuous, generic and customizable RAI KM aligned with information on lifecycle processes and design decision-making that result in the respective AI system. The process to define RAI approaches includes the identification of qualitative methods that are suitable for compliant lifecycle design.

Quality Criteria The EU AI Office,[1] which is tasked with implementing the AI Act, communicated criteria towards AI Act alignment of acceptable standards

[1] https://digital-strategy.ec.europa.eu/en/policies/ai-office

for high-risk applications [SG24b] [Eura]. These criteria function as quality-check, while exploring the integration of suitable methods, and a comprehensive coverage of regulatory requirements along the AI lifecycle is required. The following list presents an overview of required quality criteria for acceptable standards [Eura] [SG24b]:

1. The main prerequisite is a shared definition of *Risk*, which is specified as "[...] the combination of the probability of an occurrence of harm and the severity of that harm" in Article 3 in the AI Act [Fut24], and harm focuses on human beings and the planet, in contrast to harm to the organization. [Eura] Overall, compliant AI standards need to be "[t]ailored to the objectives of the AI Act" [SG24b, 3].

2. Important considerations include aligning acceptable methods with AI trustworthiness criteria [Eur19] [Hig20], and their interplay.

3. Qualitative methods need to be AI system- and product-oriented, and "[...] cover all the phases of the product lifecycle, from initial inception of the AI system, when risks can already start to be identified and assessed, up to post-market placement stages, i.e. those in which AI systems are monitored while in operation, and possibly updated and modified" [SG24b, 3].

4. The generated content needs to be sufficiently prescriptive with clear requirements for application, focusing on "[...] the criteria and priorities that AI providers must observe when implementing them, as well as when assessing compliance" [SG24b, 3].

5. Acceptable methods should be aligned with state of the art AI techniques (if existent), which necessitates continuous knowledge updates.

6. AI Act-conform-methods need to be applicable across sectors, and types of AI systems, while "[i]t should be reasonably clear for AI providers, when presented with horizontal requirements, how to identify suitable ways to apply them to their AI products in light of their intended use and the identified risks. Therefore, standards are also expected to provide the necessary guidance to support their application in specific contexts" [SG24b, 3].

In summary, a methodology towards RAI should include methods that share the same definition of risk, address AI system and product-related information, formulate clear guidelines for application, address AI's evolving state of the art, while being aligned with TAI, as well as close the gap between generalizability of selected technical approaches and individual application scenarios. In light of existing differences among domains, for instance, the medical domain in particular is characterized by data scarcity, which does not necessarily apply to other sectors to the same extent, the proposed RAI methodology is envisioned to provide sufficient customizability for fine-tuning to particular sectors, while we focus on healthcare as a start for high-risk systems.

RAI Information Management Overall, implementing RAI systems has to fulfill a multitude of intricate and partly abstract criteria centered around a complex technology that is incomprehensible to humans. Consequently, a comprehensive IM can help to shed light on underlying processes, rendering quality more tangible to stakeholders, which enables taking responsibility. For this endeavor to be successful, relevant information needs to be defined, organized along AI lifecycle processes for execution, as well as represented in an understandable manner, which is adapted to different stakeholders that appear at different stages of the lifecycle. TAI [Eur19], and the EU AI Act [Fut24] provide guidelines how we, as a society would like AI to behave, as well as concrete requirements for providers such as data documentation and a required instruction-of-use for high-risk AI. The question arises how to identify and evaluate the AI project-specific content and realization of these requirements in a proven reliable manner. The next crucial step is implementing them for individual AI systems, a challenging task, which requires valid technical guidelines towards implemented methods whose application can be proven as justified. A unified RAI information format that addresses AI's inherent dynamics along the AI lifecycle, as well as supplementary contextual RAI information contributes to the implementation of RAI systems. Focusing on the extraction of existing interdependencies and avoidance of AI pitfalls, this includes related input and output information for concrete design decision-making towards quality by design. For instance, the data distribution impacts the applicability of certain metrics, and preprocessing steps the performance of XAI methods, or evaluation on the test set is not conclusive if data normalization is executed before data splitting. Further, RM-related information should be extracted and linked with implementation, as well as the transformation of content for different stakeholder views considered. In addition, since a lot of supplementary information to close the gap to the intelligent system's real-world integration is required, which differs for individual use cases, customizability of information extraction is a desirable criteria. As a result, a methodology towards RAI should link lifecycle design decision-making with contextually relevant information in a comprehensive, generic and customizable way.

AI Design Knowledge Alignment The technology's evolutionary character and on-going research on general, as well as use case-specific design knowledge results in necessary, continuous alignment of the RAI methodology with existing publications, and novel contributions alike. Existing and evolving knowledge on how to implement RAI can be organized according to groups of use cases based on structural similarities from a technical viewpoint with respect to all AI applications, towards closing the gap between generalizability and use case-specificity. Another interesting analysis to structure design decision-making methods includes possibly differing conditions of individual domains such as education, avionics or medicine. Organizing AI lifecycle processes, and design decision-making according to real-world use cases could be a valid additional consideration. For instance, with respect to the medical domain, where the medical user plays a crucial role, knowledge how to explain AI output could be transferrable to AI in education, where students and teachers play an equally significant role. Further, knowledge on how to address AI risks in a continuous manner needs to be gathered, organized, and updated tailored to individual use cases, which necessitates a dynamic AI risk classification as a first step. This enables the consideration of relevant risks for the implementation of risk controls, which need to be adapted to specific application scenarios.

1.3.1.3. Embedded Ethics & Risk Mitigation

AI ethics and risks are closely related, the latter menace the implementation of TAI, and RM can be interpreted as the process that implements TAI.

Trustworthy AI The EU AI Act's ethical foundation [Eur19] introduces seven fundamental requirements on TAI that are derived from ethical questions surrounding fundamental rights, the planet, health, and safety, as previously outlined. They summarize general ethical criteria, and provide a structured overview tailored to AI. The seven TAI criteria equally provide guided questions in form of an *Assessment List for Trustworthy AI* (ALTAI) [Hig20] that aim to close the gap towards a practical implementation of AI, or required risk controls. This generic structure can be used as foundation for RM, as equally proposed by the American National Institute of Standards and Technology (NIST) [Tab23].

Risk Management: Highlighting the central role of RM, the RAI methodology should include addressing AI risks by design. This could start with an organized risk classification of known AI risks for a comprehensive identification of use case-specific risks, based on AI system context-related information. Next, risks are evaluated with respect to their identified probability of occurrence and severity if they are acceptable or not, which includes residual risks that emerge or stay after risk control measures are implemented. [ISO19] It is worth highlighting the iterative character of these processes, which, addressing AI's evolutionary character, continues until the system's decommissioning, and a RAI methodology should enable continuous risk monitoring.

Risk Mitigation by Design Promoting a multi-dimensional approach towards risk mitigation by design, complex AI risks need to be addressed from multiple angles, aiming for continuity, which includes the consideration of use case-adaptation of ethical questions. Implemented risk controls in form of design decisions along the AI lifecycle need to be documented, linked with identified risks, and continuously monitored as part of the previously introduced comprehensive RAI IM structure that links risks with the implemented lifecycle. Further, a unifying design decision information format along the AI lifecycle that incorporates interdependencies contributes to mitigate risks that emerge based on an incorrect implementation, for instance through information extraction, and transformation, which can be based on guiding questions. Finally, comprehensive domain embedding and alignment with the system's intended use is crucial, since the true success of the intelligent system can only be measured in the real-world. As a result, the proposed RAI methodology should incorporate risk mitigation from different angles starting from the beginning of its design so that it is created in a trustworthy manner.

Embedded Ethics The human influence, which is biased and incomplete, poses risks in itself. Humans comprise a crucial component of the AI lifecycle, which is created and maintained through their decision-making. Consequently, relevant ethical considerations need to be anticipated, and discussed by all stakeholders, as well as tailored to individual RAI projects. This necessitates interdisciplinary collaboration, the promotion of AI literacy, as well as continuous ethics training integrated from the start of RAI project planning. These factors result in a need for innovative approaches centered around "Embedded Ethics and Social Sciences" [MS20] to integrate ethics throughout the lifecycle process. Highlighting the importance to consider the human mind leads to the requirement for a RAI methodology to provide room for embedded ethics, for instance in form of ethics training by design.

1.3.2. Generic & Customizable Methodology Design

The creation of a methodology that aims to integrate AI's inherent dynamics, highlighting the technology's evolutionary character, as well as our current level of (rudimentary) RAI knowledge on implementation and impact-assessment in the real-world is an immense task. This complexity advocates for a decentralized approach to the holistic methodology's creation, as knowledge on AI risks and AI design principles continues to evolve, highlighting the need for generalizability and customizability. Long-term, these principles are envisioned to be implemented as a practical and trustworthy tool to be used during AI projects. This section introduces the abstraction of underlying processes for methodology design, highlighting necessary steps.

Essential RAI Information To develop a robust methodology for RAI, the first step involves identifying critical information blocks. These blocks encapsulate key information areas, centered around AI lifecycle phases, risk categories, and mitigation strategies, ensuring the methodology aligns with diverse use case requirements to foster broad applicability towards quality by design.

Methodological Building Blocks The methodology must incorporate building blocks that format and link information effectively. These blocks should facilitate seamless integration and traceability during AI projects, highlighting interdependency analysis, which enables a structured approach to organizing and connecting RAI-related information. Information fusion can be achieved through a generalizable information format and linking structure that closes the gap between compliance-related concepts and AI-specific dynamics. Also, the format should enable information connection with already existing methods that collect relevant information to foster a seamless real-world integration.

Generalizability & Customizability A generalizable blueprint should be developed to standardize the methodology across AI use cases. This information structure should outline an AI project application-oriented lifecycle model and address associated risks, offering a flexible blueprint, that is adaptable to varied AI scenarios. For instance, different templates that are derived from the generic blueprint provide more fixed use case-specific design guidelines. Further, the blueprint should be extendable with additional, as relevant identified RAI information blocks, aiming to address the continuous evolution of potentially necessary RAI knowledge.

Methodology Design Processes Processes involved in populating the basic blueprint structure with more specific information should be examined in detail to ensure practical applicability, and continuous refinement of the RAI methodology. Therefore, the holistic methodology must be punctually illustrated with use case-specific examples. These examples demonstrate information linking, organization, and application from multiple perspectives, highlighting reliable design decision-making, and RM. By abstracting the identified processes, populating the blueprint with design knowledge supports a decentralized and customizable approach. This setup leverages collective expertise and enables the incorporation of continuously evolving design knowledge, while simultaneously supporting AI QM during individual projects that equally apply, generate and update RAI knowledge.

1.3.3. Customizable to Different Domains – Illustrated for Medicine

An important consideration that influences the methodology's design process comprises evaluating its applicability to a specific domain, and requirements of individual sectors may differ. Medical intelligent applications are centered around human lives, which results in a lot of complex ethical questions and this perspective is particularly compelling as a starting point to design the RAI methodology for high-risk AI applications. We highlight domain-specific conditions, as well as sector-specific regulations that apply, and both impact lifecycle design choices.

Domain refers to a specific area of knowledge, activity, or expertise [DL22a, 2]. In the context of science, technology, or AI, a domain is the field or subject matter to which a system or application is applied, such as healthcare, finance, or law. Understanding the domain is important because it influences how problems are framed, which data is relevant, and how solutions should be designed for a seamless real-world integration. A domain model "[...] describes what can happen (behavior) and what can exist (state and structure) in a domain, or in other words, what can be controlled and managed in a domain" [DL22a, 2]. It "[...] is the foundation for a shared lexicon in communication between stakeholders, and can serve directly as a structured vocabulary for making other specifications [...]" [DL22a, 2].

Domain-specific Conditions The medical sector is a high-risk domain with specific challenges, such as high data scarcity and labeling costs, complex domain knowledge and privacy rights, as well as the need for interpretable explanations of AI output, tailored to the medical user or patient. Section 5.1 further explores challenges for AI in healthcare. Overall, these conditions contribute to the methodology's design, namely the four fundamental design principles, as introduced in more detail in section 3.3.2.

Sector is an economic term [BG99]. It refers to a distinct part of the economy or society, which "[...] may be divided into groups of industries so that each sector exhibits significantly different characteristics" [Wol55, 402]. Examples include the healthcare sector, education sector, or public sector. In contrast to *domain*, which often relates to knowledge or subject areas, *sector* emphasizes organizational or economic divisions and is often used in policy, business, or industry contexts.

Sector-specific Regulation Generally, healthcare counts among the highest regulated sectors, and the AI Act poses additional regulatory requirements for high-risk intelligent systems on a horizontal level, in addition to existing sector-specific, or vertical regulation, i.e. the Medical Device Regulation (MDR) (or In Vitro Diagnostic Medical Device Regulation (IVDR)). From a legal viewpoint, as outlined in Annex I, the AI Act and MDR/IVDR are harmonized [Fut24], and "[f]or algorithms embedded in products where sector regulations apply, such as medical devices, the requirements stipulated in the [AI Act] will simply be in-corporated into existing sectoral testing and certification procedures" [FL22, 3]. A legal analysis of the MDR/IVDR and AI Act is evaluated in more detail in [WC22], and "[b]y means of the concept that the more product specific regulation applies, but more specific health and safety requirements must be met under horizontal legislation, it is ensured that devices under the MDR and IVDR are not subjected to a double regulatory burden. It is the intention of the Commission that the AI Act will have this effect" [WC22, 2].

Sector-specific Quality Management With respect to the implementation of legal requirements, standards play a crucial role and, focusing on the medical domain, a structured identification and management of critical risk factors contributing to QM already exists. Consequently, "it is expected that medical device manufacturers will continue using the ISO/IEC 13485 QMS and incorporate the requirements of Art. 17 within their existing medical device QMS functions" [AM24a, 2]. For instance, the MDR includes prerelease assessment and a post-market monitoring structure, which is equally demanded by the AI Act for high-risk AI. This results in the need for the RAI methodology to enable the integration of identified information blocks for AI-specific QM procedures with existing QM structures. Relevant stakeholders that participate in AI QM include notified bodies who execute compliance assessment, and providers of AI systems alike. Finally, adding complex horizontal requirements to already prevailing challenges regarding the concrete implementation of regulatory requirements imposed by the vertical regulation for individual use cases might result in slowing innovation [AM24a, 4]. This necessitates the development of tools that support the compliance assessment process in a reliable manner.

1.3.4. Evaluation through Use Cases

As a final step for methodology design, its applicability needs to be evaluated based on use cases to test the proposed setup focusing on information extraction, transformation and fusion. The selected use cases and application scenarios should each focus on different components that together offer a good estimation of the fundamental generic structure. Overall, aiming for a methodology towards RAI that is sufficiently complete, while providing the desired customizability for application.

Diverse Medical Background Focusing on a medical application scenario, selected fictional or non-fictional use cases should represent diverse medical specialties and usage scenarios [ISO21b], while being situated within the same lifecycle phase. This approach not only allows the phase's definition to be highlighted from multiple perspectives but also ensures that the extracted workflows are transferable to other stages of the lifecycle, including the extraction of relevant information throughout the complete lifecycle. Additionally, the inclusion of relevant medical knowledge is essential to align the methodology's design with supplementary contextual information blocks, illustrating a generalizability of information extraction for the nuances of each medical field.

Different Perspectives on the Design Process Use cases to test the methodology should showcase diverse perspectives of the methodology's design process, facilitating the abstraction of the design workflow. This approach aims to support the decentralized continuation of developing and applying the methodology, ensuring adaptability and scalability across varied contexts. The design process emphasizes defining AI lifecycle processes from an application-oriented perspective, while ensuring seamless integration with supplementary contextual information, such as RM frameworks, or the incorporation of stakeholder perspectives.

Reliable Design Decision Making & Building Blocks Given the current state of AI knowledge, use cases play a critical role in identifying reliable lifecycle process execution and informed design decision-making. They support the identification and application of the methodology’s foundational building blocks, including the implementation and extraction of lifecycle design guidelines and their integration with supplementary RAI knowledge. Focusing on the medical domain, the evaluation of use cases should demonstrate how the methodology addresses diverse stakeholder needs, incorporates domain-specific knowledge, and analyzes technical interdependencies, ensuring alignment with both ethical principles and practical objectives.

1.4. *Design Science Research*

To design the RAI methodology, we follow principles from *Design Science Research* (DSR) [vJ20]. As summarized in Figure 1.1, DSR describes the on-going and dynamic communication between design knowledge and application of that knowledge. DSR is introduced in more detail in section 3.3.7.

Incorporating DSR, we focus on the continuous interplay between an abstract knowledge base and use case-derived findings within a specific environment to incorporate adaptability and testing, aiming to derive, order, and assign rules for prospective AI lifecycle planning towards risk mitigation by design: Based on the creation of *DSR Artifacts*, a contribution to *Design Knowledge* as a whole is made: "DSR aims to generate knowledge of how things can and should be constructed or arranged (i.e., designed), usually by human agency, to achieve a desired set of goals; referred to as design knowledge" [vJ20, 2].

The concrete processes to define the interplay between abstract design knowledge and concrete use cases, once identified, can happen in a decentralized manner, which depends on a critical mass of participants. We aim to develop a blueprint that allows the fusion of findings integrated along AI lifecycle stages in an organized manner to provide suitable approaches for AI use cases through a unified information format that incorporates RAI knowledge.

The Generic and Customizable Methodology based on Quality Gates towards Certifiable AI in Medicine (MQG4AI), as presented in this thesis, including the outlined building blocks and structure, is designed based on DSR, through a retrospective analysis of our medical and technical use cases, each of which serves to highlight different components.

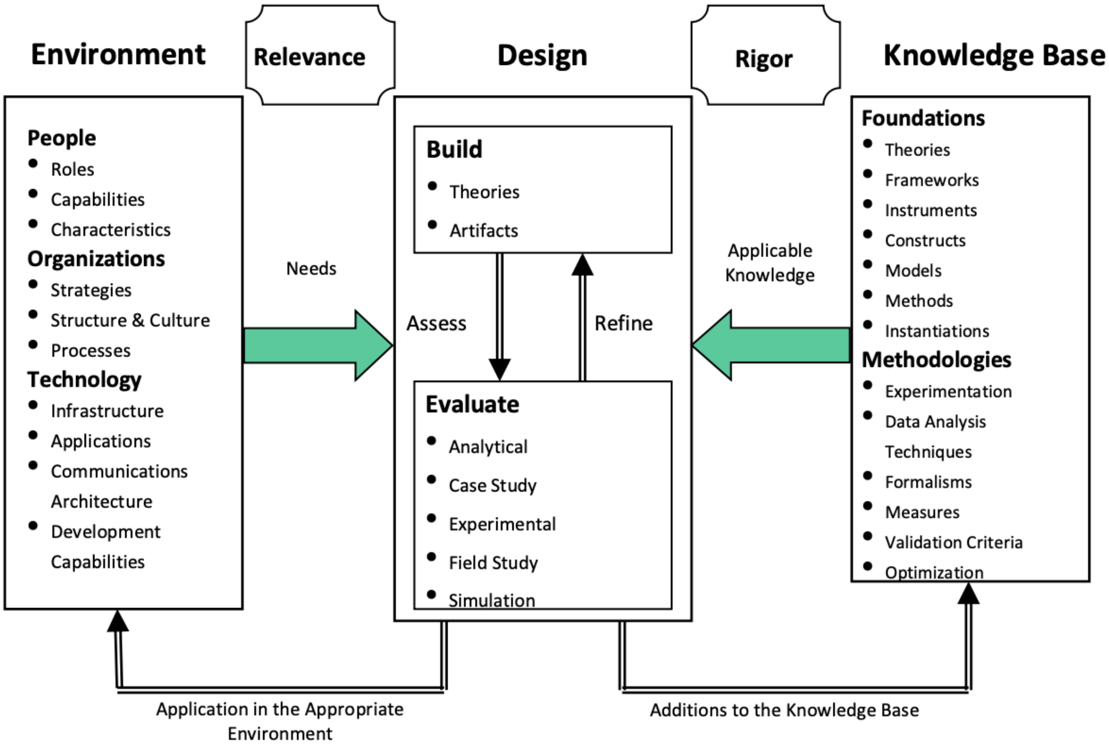


Figure 1.1.: Design Science Research (DSR) processes, inserted from [vJ20, 4]

1.5. Thesis Outline

Generally, this manuscript is to be understood as a snapshot, laying the foundation for a *living document* that aims to establish a dynamic blueprint in the form of MQG4AI, as depicted in Figure 1.2, towards implementing RAI. Therefore, the introduced concepts and building blocks, ethical foundation, as well as RAI information collections are designed to incorporate evolving and continuous dynamics. This equally aligns with AI systems that are based on a stochastic and opaque technology.

Depending on the reader's prior knowledge, this work is structured to be read linearly, with each chapter building on the previous one. Part I lays the foundational knowledge regarding Deep Neural Networks (DNN) as a technology. Part II introduces the concept of MQG4AI, including its context surrounding RAI. Part III focuses on the evaluation of MQG4AI, drawing from the information provided in earlier chapters to illustrate how to interact with MQG4AI. Chapter 8 may be read independently, while chapters 6 and 7 should be read linearly. Finally, part IV is designed to be read stand-alone, offering a summarized overview of our contribution, as well as future work. This part can also be read in combination with the introduction for a more concise understanding of our key proposition. How the chapters relate in the context of MQG4AI is depicted in Figure 1.2.

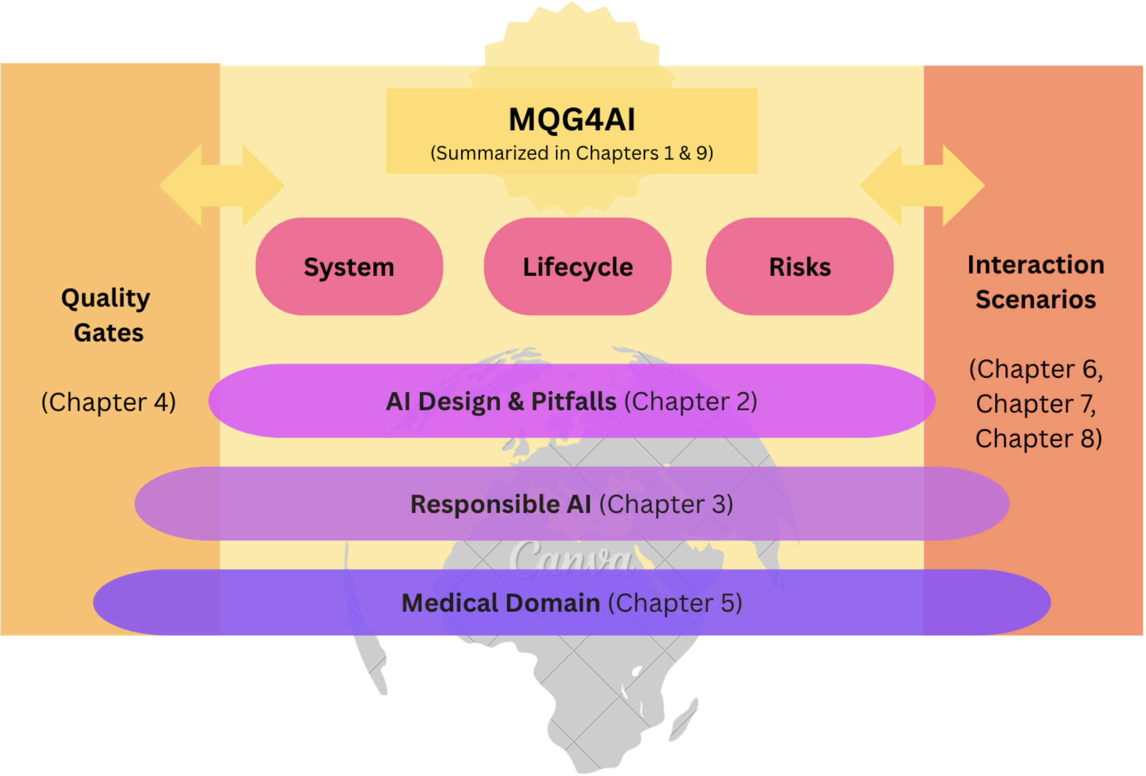

MQG4AI
(Summarized in Chapters 1 & 9)
Quality
Gates
(Chapter 4)
System
Lifecycle
Risks
Interaction
Scenarios
(Chapter 6,
Chapter 7,
Chapter 8)
AI Design & Pitfalls (Chapter 2)
Responsible AI (Chapter 3)
Medical Domain (Chapter 5)
Canva

Figure 1.2.: A visualization of how the chapters of this thesis are related within the context of MQG4AI. Chapters 2, 3, and 5 provide foundational knowledge and introduce the concept. Chapter 4 outlines MQG4AI's core building blocks and chapters 6, 7, and 8 simulate MQG4AI interaction scenarios for evaluation and illustration of applicability. Chapters 1 and 9 motivate and summarize the core idea, including future work.

Part I. Introduction and Foundations

Chapter 1: Introduction Chapter 1 introduces the motivation and vision behind this thesis, framed by the EU AI Act and RAI, especially in sensitive domains like medicine. It outlines key challenges such as the complexity of DNNs, the lack of standardized AI Quality Management (QM), and the ethical implications of AI use. Next, the chapter defines research objectives, presents our research question, and introduces our method *Design Science Research* (DSR) [vJ20]. Finally, contributions through projects, publications, and AI-assisted writing are highlighted.

Chapter 2: Foundations Chapter 2 provides the foundational knowledge necessary for understanding and supporting the creation of DNNs, with a focus on their design. It highlights key aspects that influence design decision-making, including the type of AI system, data characteristics, domain knowledge, mathematical limitations and computing power, evaluation metrics, and explanation techniques. By examining these interrelated factors, the chapter identifies critical design considerations and common pitfalls, aiming to introduce the intricacies of DNNs from a development perspective.

Part II. Methodology based on Quality Gates towards Certifiable AI in Medicine (MQG4AI)

Chapter 3: Context, Method and Objective Chapter 3 sets the stage for the MQG4AI concept by introducing the broader context of RAI, framed by its core dimensions: lawfulness, ethics, and accountability [DRN23]. It outlines the structure of the AI lifecycle and highlights the central role of Risk Management (RM) as a foundation for quality-related processes. Building on this, the MQG4AI blueprint is introduced as a lifecycle planning method, centered around structured information and knowledge flows. The supplementary information blocks that support reliable, dynamic lifecycle design are introduced. Next, the chapter details the integration with DSR [vJ20] in form of project-specific application (MQG4A) and RAI design knowledge (MQG4DK), that communicate with each other. Finally, vision and limitations of the proposal are outlined, with a focus on advancing AI QM in practice.

Chapter 4: Setup, Components and Structure Chapter 4 outlines the central building block of MQG4AI: Quality Gate (QGs). It begins by presenting a generic, high-level AI lifecycle structure to anchor the concept of QGs across all phases, introducing *collection-QGs*. The chapter then introduces the detailed *leaf-QG* information template, which enables systematic integration of design decision-relevant information. Finally, key dimensions for knowledge fusion within MQG4AI (MQG4A and MQG4DK) are discussed, with a focus on organizing, contextualizing, and tracing decisions critical to AI design. Together, these elements operationalize quality-oriented lifecycle planning and foster transparency in development workflows.

Chapter 5: Application Domain Medicine Chapter 5 introduces the medical domain as a high-stakes context for AI applications, emphasizing its complex regulatory landscape, ethical requirements, and challenges related to medical data quality and availability, among others. These domain-specific factors critically influence responsible and high-quality AI design. The chapter then presents two use cases used for evaluating MQG4AI in part III: a fictional multi-label classification task based on ElectroCardioGram (ECG) data and a real-world case addressing the rare disease Achalasia with intelligent and digital enhancements within the prototypical software *EsophagusVisualization*.

Part III. Application and Evaluation

Chapter 6: MQG4DK Design – Defining the *Explanation* Stage during Model Development & Introduction of the *Leaf-QG* Chapter 6 proposes a generic and comprehensive definition of the explanation stage as a critical phase in the AI lifecycle, guided by risk-awareness and best practices to foster quality by design [EM25a]. The stage is aligned with emerging standardization efforts [Art24] and emphasizes the integration of explainability into RAI development. Furthermore, the chapter presents a concrete example of a Leaf-QG contribution to MQG4DK: a technical guideline for evaluating the robustness and fidelity of explanations [LGC24].

Chapter 7: MQG4DK Design – *Leaf-QG* Compilation for Reliable Performance Evaluation Metrics based on a Fictional Use Case Chapter 7 presents a leaf-QG contribution to MQG4DK that consists of a compilation of leaf-QGs, focusing on the reliable performance evaluation of AI systems. Building on the previously introduced fictional ECG multi-label classification use case, the chapter operationalizes performance-related design decisions by structuring relevant information, metrics, and classification evaluation practices in alignment with supplementary information. This contribution exemplifies how MQG4AI can guide context-sensitive and lifecycle-anchored QM. The contribution is grounded in empirical experiments [EM24] [EM25c], providing practical insights into the evaluation of model behavior in dynamic, multi-label classification scenarios.

Chapter 8: MQG4A template Versions – QG *Model Configuration* for Timed Barium Esophagogram Image Segmentation with a Human-in-the-Loop Chapter 8 focuses on MQG4A, where we conduct a retrospective analysis of a segmentation model use case within the previously introduced *EsophagusVisualization* tool. We illustrate how various related MQG4A template versions contribute to design decision-making, focusing on model configuration. Starting with a basic version that provides relevant supplementary information, we demonstrate how different model configuration versions are derived. Each version serves as a distinct lifecycle implementation approach, represented as a selection of QGs. This process highlights the flexibility and adaptability of MQG4AI in supporting decision-making throughout the AI lifecycle. The experiments, which are not the main focus, are based on a supervised bachelor thesis.

Part IV. Conclusion and Outlook

Chapter 9: MQG4AI Roadmap Chapter 9 concludes our contribution and serves as a stand-alone summary of this thesis. We first summarize MQG4AI with respect to our research objectives, including its context within RAI, its key building blocks, and its relevance to healthcare. The chapter then transitions into a reflection on MQG4AI as a tool, highlighting the relevance of on-going dynamic blueprint design, as well as flexible implementation of MQG4AI's building blocks, and concludes by outlining potential future work.

1.6. Scientific Participation

This chapter offers a summary of key contributions and experiences that that have shaped and enriched the work presented in this thesis. First, we introduce related publications and conferences, before we highlight interdisciplinary projects, with a strong focus on healthcare and Responsible AI (RAI). Finally, supervised thesis are outlined.

1.6.1. Publications

This section introduces selected events and publications that have contributed to the development of this thesis. First, we highlight publications that are directly related to this thesis. Next, we introduce further, more broadly RAI-related publications, including two upcoming projects. Finally, we highlight a combination of events that indirectly contributed to the creation of this thesis.

Passages and findings relevant to this thesis, and previously published, are incorporated into this thesis. They are not always cited.

Responsible AI, ethics, and the AI lifecycle: how to consider the human influence? *Miriam Elia, Paula Ziethmann, Julia Krumme, Kerstin Schlögl-Flierl & Bernhard Bauer*, in *Springer – AI and Ethics 2025*[2]

(Abstract) Continuing the digital revolution, AI is capable to transform our world. Thanks to its novelty, we can define how we, as a society, envision this fascinating technology to integrate with existing processes. The EU AI Act follows a risk-based approach, and we argue that addressing the human influence, which poses risks along the AI lifecycle is crucial to ensure the desired quality of the model's transition from research to reality. Therefore, we propose a holistic approach that aims to continuously guide the involved stakeholders' mindset, namely developers and domain experts, among others towards Responsible AI (RAI) lifecycle management. Focusing on the development view with regard to regulation, our proposed four pillars comprise the well-known concepts of Generalizability, Adaptability and Translationality. In addition, we introduce Transversality (Welsch in Vernunft: Die Zeitgenössische Vernunftkritik Und Das Konzept der Transversalen Vernunft, Suhrkamp, Frankfurt am Main, 1995), aiming to capture the multifaceted concept of bias, and base the four pillars on Education, and Research. Overall, we aim to provide an application-oriented summary of RAI. Our goal is to distill RAI-related principles into a concise set of concepts that emphasize implementation quality. Concluding, we introduce the ethical foundation's transition to an applicable ethos for RAI projects as part of on-going research.

[2] https://link.springer.com/article/10.1007/s43681-025-00666-z

This journal paper emphasizes the necessity of integrating ethics throughout the AI lifecycle and demonstrates how this integration can be operationalized using the proposed approach, which is also implemented and documented in the MQG4AI visualization available on GitHub[3], as further detailed in Part II of this thesis. The initial idea emerged from an exploratory dialogue with Chat-GPT, which unexpectedly evolved into a philosophical exchange. During this interaction, the AI introduced the term *transversal*. The prompts consisted of asking if terminology existed for concepts and methods that aim to summarize the complexities of the real world towards unity. This revelation prompted further reflection on the term's applicability to ethics integration along the AI lifecycle. To validate and refine the concept, ethical domain experts involved with the CReAITech [JS23] were consulted, leading to a constructive interdisciplinary exchange that helped shape the overarching blueprint presented in this thesis. In pursuit of practical applicability, the approach was further developed in collaboration with experts of the DARE-method [KJ24], who contributed exemplary materials supporting the structured integration of ethical reflection throughout project-based AI development.

MQG4AI Towards Responsible High-risk AI - Illustrated for Transparency Focusing on Explainability Techniques *Miriam Elia, Alba Maria Lopez, Katherin Alexandra Corredor, Bernhard Bauer, Esteban Garcia-Cuesta*, as *pre-print on arXiv 2025*[4]

(Abstract) As artificial intelligence (AI) systems become increasingly integrated into critical domains, ensuring their responsible design and continuous development is imperative. Effective AI quality management (QM) requires tools and methodologies that address the complexities of the AI lifecycle. In this paper, we propose an approach for AI lifecycle planning that bridges the gap between generic guidelines and use case-specific requirements (MQG4AI). Our work aims to contribute to the development of practical tools for implementing Responsible AI (RAI) by aligning lifecycle planning with technical, ethical and regulatory demands. Central to our approach is the introduction of a flexible and customizable Methodology based on Quality Gates, whose building blocks incorporate RAI knowledge through information linking along the AI lifecycle in a continuous manner, addressing AI's evolutionary character. For our present contribution, we put a particular emphasis on the Explanation stage during model development, and illustrate how to align a guideline to evaluate the quality of explanations with MQG4AI, contributing to overall Transparency.

This pre-print summarizes key aspects of this thesis, with a specific focus on the Explanation stage of the AI lifecycle, as introduced in Chapters 4 and 6. The paper was developed in collaboration with experts in explainability and RAI from

[3] https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/1_System/Ethics_General/Ethics_General.md

[4] https://arxiv.org/abs/2502.11889

the Universidad Politécnica de Madrid (UPM). Beyond contributing to the conceptualization of the explanation lifecycle stage, their professional insights significantly informed the development of the supplementary information blocks within the MQG4AI methodology, as well as the proposed information flow and other building blocks, such as template versioning. The proposed lifecycle planning blueprint is detailed in chapter 3. Although the initial journal submission was not accepted, our team is currently preparing a revised manuscript that presents the core ideas more accessibly and with a clearer narrative. The updated storyline and structure of the follow-up publication will be introduced later.[5]

Towards Certifiable AI in Medicine: Illustrated for Multi-label ECG Classification Performance Metrics *Miriam Elia, Fabian Stieler, Fabian Ripke, Marius Nann, Sarah Dopfer, Bernhard Bauer*, at the *2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS) in Madrid*[6]

(Abstract) Cardiovascular diseases count among the most critical and propagated diseases worldwide. Artificial Intelligence [AI] generates promising results across various medical domains and can enhance cardiovascular diagnosis and treatment. However, contextual knowledge is crucial for multiple design decisions to avoid pitfalls, and the true success of the intelligent application can only be measured in relation to its intended clinical setting. Machine learning [ML] models are evaluated based on performance metrics whose interpretation is influenced by data distribution and tuning objective. The present paper introduces a domain-embedded approach aiming towards a reliable performance evaluation of ML models throughout the complete lifecycle. Our findings are illustrated for multi-label ECG classification in an emergency setting and calculated on open-source Physionet data. Finally, the resulting procedure is embedded as a contribution to Quality Gate Metrics within our generic and customizable methodology based on Quality Gates towards certifiable AI in medicine.

This conference paper forms the technical foundation of the experimental work that substantially contributed to the development of this thesis. Developed in collaboration with the medical device manufacturer corpuls and embedded within the LIFEDATA pipeline, the study leveraged their advanced infrastructure, including a parallelizable development and production environment as well as features for AL [SF23]. The research involved a retrospective analysis of design decisions along the pipeline, focusing on establishing a reliable design approach for the development stage within the AI lifecycle. Experimental materials are provided on Zenodo[7]. While the main emphasis was placed on performance metrics, additional components such as loss functions and explainability algorithms were

[5] Note that the pre-print was published before we introduced *(living) blueprint* to describe MQG4AI, from previously solely referring to the entire concept as *template*. The latter is now applied to the MQG4A context, while the former evolves into MQG4DK.

[6] `https://ieeexplore.ieee.org/document/10570023/authors#authors`

[7] `https://zenodo.org/records/14652465`

also analyzed in depth, particularly in the context of extreme class imbalance. For example, *sign loss* [ZZ21] in the context of ECGs was experimented with, as well as *focal loss* [Lin18], which is commonly used in object detection due to its ability to manage background-heavy data. Both methods propose to differently "punish" misclassification of prevalent versus rare samples. Among others, these practical insights informed the derivation of the structured pre-, intra-, post-selection steps for organizing information throughout the development workflow, which we propose to be realized through different lifecycle planning template versions, as introduced in chapter 8. Additionally, in collaboration with domain experts, a fictional use case situated in emergency medicine was developed, as introduced in the section 5.2.1. This use case serves as the basis for the analytical discussion presented in chapter 7, underscoring the importance of domain-specific considerations in AI design decision-making.

A Methodology Based on Quality Gates for Certifiable AI in Medicine: Towards a Reliable Application of Metrics in Machine Learning *Miriam Elia and Bernhard Bauer*, at the *18th International Conference on Software Technologies 2023 (ICSOFT) in Rome*[8]

(Abstract) As of now, intelligent technologies experience a rapid growth. For a reliable adoption of those new and powerful systems into day-to-day life, especially with respect to high-risk settings such as medicine, technical means to realize legal requirements correctly, are indispensible. Our proposed methodology comprises an approach to translate such partly more abstract concepts into concrete instructions - it is based on Quality Gates along the intelligent system's complete life cycle, which are composed of use-case adapted Criteria that need to be addressed with respect to certification. Also, the underlying philosophy regarding stakeholder inclusion, domain embedding and risk analysis is illustrated. In the present paper, the Quality Gate Metrics is outlined for the application of machine learning performance metrics focused on binary classification.

This conference paper marks the initial presentation of the core ideas developed in this thesis. Winning the *Best Position Paper Award* definitely motivated on-going research. The paper outlines the fundamental building blocks of the proposed methodology and introduces an early version of our AI lifecycle process view. In doing so, it also addresses key challenges in designing AI lifecycles that are compliant with regulatory and quality standards, using the example of reliable performance evaluation metrics. The concepts introduced in this paper laid the groundwork for the more refined version of lifecycle design and building blocks presented in the final version of this thesis, as detailed in chapter 4.

Further Publications In addition to the core publications directly integrated into the structure of this thesis, several further publications address topics closely related to the thesis's overarching themes, namely RAI in medicine. These works

[8] `https://www.scitepress.org/Link.aspx?doi=10.5220/0012121300003538`

explore foundational and complementary aspects such as AI ethics in medicine, ethical integration, quality assessment, and design processes within real-world AI development. Together, they reflect the broader research context and provide valuable insights that have informed and enriched the development of this thesis.

- *Ziethmann, P., Elia, M., Stieler, F. et al. Clinical Decision Support Systems at the Intersection of Technology and Ethics: A Critical Analysis of the Ethical Guidelines Issued by the German Medical Association. Digit. Soc. 4, 15 (2025).* `https://doi.org/10.1007/s44206-025-00175-w`

- *Ziethmann, P., Stieler, F., Elia, M., et al. CDSS – An Interdisciplinary Perspective on the Statement of the Central Ethics Commission of the German Medical Association, The Royal College of Radiologists Open, 3, 1 (2025).* `https://doi.org/10.1016/j.rcro.2024.100163`

- *Grünherz, V., Ebigbo, A., Elia, M., et al. Automatic three-dimensional reconstruction of the oesophagus in achalasia patients undergoing POEM: an innovative approach for evaluating treatment outcomes, BMJ Open Gastroenterology (2024).* `https://doi.org/10.1136/bmjgast-2024-001396`

- *Boulogne, H.L., Lorenz, J., Kienzle, D., et al. The STOIC2021 COVID-19 AI challenge: Applying reusable training methodologies to private data, Medical Image Analysis, 97 (2024)* `https://doi.org/10.1016/j.media.2024.103230.`

- *Nagl, S., Grünherz, V., Elia, M., et al. Automatic Three-Dimensional Reconstruction of the Esophagus in Achalasia Patients undergoing POEM: a Comprehensive Assessment of Treatment Outcomes and pathophysiological Changes, Endoscopy, Georg Thieme Verlag KG (2024).* `https://doi.org/10.1055/s-0044-1783095`

- *Stieler, F., Elia, M., Weigell, B., et al. LIFEDATA - A Framework for Traceable Active Learning Projects, IEEE 31st International Requirements Engineering Conference Workshops (REW), Hannover, Germany, pp. 465-474 (2023)* `https://doi.org/10.1109/REW57809.2023.00088`

- *Müller, D., Mertes, S., Schröter, N., et al. Towards Automated COVID-19 Presence and Severity Classification, Stud Health Technol Inform. (2023)* `https://doi.org/10.3233/SHTI230309`

- *Elia, M., Peter, T., Stieler, F., et al. Precision Medicine for Achalasia Diagnosis: A Multi-modal and interdisciplinary Approach for Training Data Generation, IEEE International Symposium on Biomedical Imaging, Cartagena de Indias, Colombia, Abstract (2023)* `https://doi.org/10.13140/RG.2.2.17880.06402`

- *Elia, M., Gajek, C., Schiendorfer, A., et al. An Interactive Web Application for Decision Tree Learning, Proceedings of the First Teaching Machine Learning and Artificial Intelligence Workshop, Proceedings of Machine Learning Research (2021)* `https://proceedings.mlr.press/v141/elia21a.html`

1.6.2. Projects & Collaborations

This section highlights projects and collaborations that enabled experiencing a diverse set of relevant stakeholders, emphasizing different professional backgrounds. This practical experience has profoundly shaped the interpretation of RAI, which strongly impacts the creation of this thesis.

Center for Responsible AI Technologies (CReAITech)

The CReAITech[9] is a joint initiative by the Technical University of Munich, the Munich School of Philosophy, and the University of Augsburg. It brings together expertise in technology, ethics, philosophy, and the social sciences to advance interdisciplinary research on the responsible development and application of AI in science and society.

Two publications originated as interdisciplinary cooperations. *Responsible AI, ethics, and the AI lifecycle: how to consider the human influence?* [EM25b] focuses on the interface between technology and ethics and how to combine both views along the AI lifecycle focusing on contributing stakeholders. The holistic approach is summarized and published as an official press release by the University of Augsburg.[10] The other related publications, *CDSS – An Interdisciplinary Perspective on the Statement of the Central Ethics Commission of the German Medical Association* [ZP25a] and *Clinical Decision Support Systems at the Intersection of Technology and Ethics: A Critical Analysis of the Ethical Guidelines Issued by the German Medical Association* [ZP25b], are situated at the interface of ethics, medicine, and technology, where we analyze the technical implementation and feasibility of ethical criteria for AI in medicine. Finally, the medical project centered around the rare disease Achalasia, which we develop as part of higher education, comprises a CReAITech use case [JS23].

AI Production Network Augsburg

The AI Production Network Augsburg[11] focuses on AI-based production tech-

[9] `https://center-responsible-ai.de/en/startseite/`

[10] `https://www.uni-augsburg.de/de/campusleben/neuigkeiten/2025/04/11/ganzheitlicher-ansatz-fur-verantwortungsvolle-ki/`

[11] `https://www.uni-augsburg.de/en/forschung/einrichtungen/institute/ki-produktionsnetzwerk/`

nologies at the intersection of materials, manufacturing, and data-driven modeling. It brings together research and industry to develop future-ready, AI-supported production solutions. Their showroom[12] serves as both a research showcase and a networking hub, offering hands-on insights into *AI in Production*. Interactive exhibits demonstrate AI applications in Industry 4.0, from emotion recognition for future workplaces, to LEGO® models explaining data circuits. The space also features info boards and digital displays highlighting the network's research areas.[13]

As part of a certification-focused project in collaboration with a company in the aviation sector, we analyzed the European Union Aviation Safety Agency's (EASA) approach to certifying production processes, placing particular emphasis on human- teaming and online learning. In addition, we explored supplementary sources related to the certification of intelligent systems to help outline a potential framework for certifying AI-supported production, namely non-destructive testing, in avionics.

Life Sciences Improved by a Framework for Efficient Data Annotation Through Active Learning (LIFEDATA)

The collaborative project was supported by the German Federal Ministry of Education and Research (BMBF)[15]. Project partners include the University of Augsburg, GS Elektromedizinische Geräte G. Stemple GmbH (Corpuls) in Kaufering, and the German Heart Center at the Technical University of Munich. The LIFEDATA project addresses the challenge of training reliable AI models in medicine, where large, well-annotated datasets, especially those including rare diagnoses, are often lacking. To tackle this, the project develops an open-source framework that combines Active Learning (AL) and DNNs to efficiently select and label the most informative data points. This reduces manual effort and enhances model performance. Two life science use cases (skin lesion segmentation and the classification of multi-label ECG data) demonstrate the framework's effectiveness across different data types. By integrating explainability and semi-supervised learning, LIFEDATA aims to support AI-driven diagnostics and ultimately improve patient care. Therefore, a central aim was to implement algorithms that enhance transparency by explaining how machine learning models make their decisions. [SF23]

Contributing to the LIFEDATA project during its final year, valuable experience was gained in AL and how to consider annotator agreements, hyperparameter tuning, and explainability, as well as close interaction with medical domain ex-

[12] `https://www.uni-augsburg.de/en/forschung/einrichtungen/institute/ki-produktionsnetzwerk/showroom/`

[13] I have participated in science communication events at the AI Production Network, particularly emphasizing collaboration with STEM education initiatives.[14], see the appendix V.

[15] Reference number 031L9196B

perts such as anesthesiologists and medical residents, in addition to collaborating in a joint team with other data scientists and machine learning engineers. The project's close ties with the industry provided equally a great opportunity to explore the regulatory landscape of medical AI. The following related publications are based on experiments using the LIFEDATA setup. In *A Methodology Based on Quality Gates for Certifiable AI in Medicine: Towards a Reliable Application of Metrics in Machine Learning* [EB23], we introduce the concept for this thesis focusing on performance evaluation metrics. The follow-up publication *Towards Certifiable AI in Medicine: Illustrated for Multilabel ECG Classification Performance Metrics* [EM24] proposes a generalizable approach for reliable performance evaluation metrics and derives experiment-based guidelines for (ECG) multi-label performance evaluation metrics, which is the foundation for our analysis in chapter 7. The evaluation is based on a fictional use case situated in emergency medicine, that is detailed in section 5.2.1.

1.6.3. Supervised Thesis

Several bachelor's and master's theses were supervised, highlighting valuable contributions to our research on the rare disease Achalasia, as detailed in section 5.2.2. In addition to this core focus, supervised theses also explored broader topics such as Generative AI (GenAI) and other emerging research questions at the intersection of RAI and medicine, allowing students to engage with cutting-edge developments while contributing meaningfully to ongoing projects. Generally, research on topics related to this thesis was accompanied by a multitude of supervised seminars, research, and project modules. The following thesis were supervised from most recent to earliest:

- *Interdisciplinary Design and Implementation of a Database centered around a Novel, Multi-modal 3D Reconstruction Data Type for Achalasia Diagnosis and Treatment to Enhance Statistical Medical Research Anticipating the Future Application of Artificial Intelligence* (Schneider, 2025) The master's thesis is situated in the field of databases. As part of our project to develop a prototype medical software to support the diagnosis, treatment, and research of the rare disease Achalasia, a motility disorder of the esophagus, a multifunctional database was designed and a prototype implemented. This includes a Graphical User Interface (GUI) extension for entering the complex clinical structure of medical procedures. Based on close collaboration with a gastroenterologist, an Entity-Relationship (ER) model was developed that reflects the clinical reality of Achalasia patients. The database can store multimodal clinical raw data, including a new data type: a 3D reconstruction of the esophagus annotated with pressure values, which forms the core of the EsophagusVisualization software. The goal of the database is to store all necessary data to ensure ongoing medical research. Furthermore, as the dataset continues to grow, it is intended to enable the application of AI scenarios. Namely, GenAI for the automated creation of 3D reconstructions during endoscopy, and Clinical Decision Support (CDSS) for personalized therapy recommendations.

- *Oesophageal Segmentation using Barlow Twins and Multi-label Annotation in the context of the Rare Disease Achalasia – A Self-Supervised Learning Approach* (Kupfer, 2025) The bachelor's thesis is situated in the field of machine learning. Specifically, a segmentation model was created as part of our project on Achalasia. This contribution focuses on optimizing the interaction workflow between medical users and the software. Using human input and multimodal data, a 3D reconstruction of the esophagus is generated. One type of input data is Timed Barium Esophagogram (TBE) images (i.e., X-ray images of the esophagus with contrast fluid). In these, the medical user outlines the shape of the esophagus as a basis for 3D reconstruction. Based on this thesis, a prototype segmentation model was integrated to enhance this step. Various approaches were selected, tested, and optimized, under the challenge of having only a small labeled dataset (85 images). The final model architecture, based on the state-of-the-art nn-Unet, was identified as the best-performing option and integrated into the tool. This analysis comprises the foundation to illustrate the, in this thesis introduced idea in chapter 8.

- *Prompt Engineering in the Context of Automated Customer Inquiry Response – Concept and Prototypical Implementation* (Koci, 2024) The bachelor's thesis involves the development of a prototype to conduct an initial evaluation of Large Language Model (LLM) integration in a medium-sized enterprise using the OpenAI API. It represents a meaningful collaboration between research and industry. In cooperation with exali AG, the performance of LLMs is assessed in the context of responding to customer inquiries within the IT insurance sector. The company supports the work with data, domain knowledge, and a multidisciplinary team of experts involved in the design and evaluation of the prototype. Within this context, a methodological framework was developed, planned, and implemented to address challenges such as limited data availability (e.g., synthetic LLM-based generation), business requirements (pricing models, data security, etc.), and evaluation approaches (technical and human-based). The goal is to provide a well-founded projection regarding the feasibility and utility of this application scenario for the company.

- *ResearchConnect - A Smart Web-Application to Enhance Interdisciplinary Work through Intelligent Matching of Scientists based on their Disciplines* (Nguyen, 2024) This bachelor's thesis explores the intriguing question of how interdisciplinary collaboration can be supported through technical means. It begins by highlighting the relevance of interdisciplinary work in research and addressing current challenges in this context. As a practical example, a prototype of an intelligent web application called *ResearchConnect* was developed based on the University of Augsburg's website. This tool enables users to identify other researchers working on similar topics via an intelligent semantic search. The system gathers information from the publicly available staff web-pages, making it easier to connect across disciplines.

- *Integration of Multi-modal Data Sources in an Application for Esophagus Analysis to Support the Diagnosis of Achalasia* (Peter, 2022) The master's thesis was conducted in collaboration with the University Hospital Augsburg (UKA) to support physicians in the diagnosis of Achalasia, kick-starting our collaborative project. Various data from several patients were provided for this purpose. Currently, doctors must manually integrate results from different data sources to achieve the most accurate diagnosis of the disease type and the resulting treatment strategy. To assist with this process, a prototypical software is developed to create the most realistic 3D reconstruction of the esophagus, which includes measured manometry pressure values as a color gradient for different measurement time points. The workflow directly includes the user during data input, utilizing manometry measurements, a barium swallow X-ray image, and images from an endoscopy for reconstruction. To integrate the data, additional information is needed, with the X-ray image, which represents the course of the esophagus, serving as the foundation. The known positions of the upper and lower esophageal sphincters are used to map the manometry measurements, and the device's measurement intervals are also known. For mapping the diameter from the endoscopic images, the measurement starting point is marked. Additionally, the height of each image must be known. Automated methods for extracting the diameter from the endoscopic images were integrated into the final software. However, due to varying lighting conditions, this extraction does not always succeed, and the software provides functionality for the user to manually adjust the selected diameter in each image. The same applies to the extraction of the esophagus course from the X-ray image. To achieve optimal usability that meets the needs of clinical users, intensive collaboration and feedback were prioritized during the software development process. The current state of our project is introduced in section 5.2.2.

This chapter introduces our proposition, research question, objectives, and vision towards implementing high-quality and responsible AI. Next, we turn to the forms and foundations of intelligent systems, exploring DL, its history and core principles, before detailing the context surrounding the creation of MQG4AI.

2 Foundations

This chapter provides a foundational overview of Artificial Intelligence (AI), outlining its fundamental concepts, methodologies, and challenges. It aims to highlight key technical considerations and related challenges in a structured manner necessary for implementing Responsible AI (RAI). The following sections are intended to reflect AI lifecycle processes, that comprise all design decisions surrounding the intelligent system's transition from idea to reality. The AI lifecycle, as well as AI risks are introduced in more detail throughout the next chapter 3. Generally, the foundations introduced in this chapter are intended to highlight intricacies surrounding AI design decision-making to motivate the benefits of a lifecycle planning blueprint (MQG4AI), which is proposed in this thesis. The following sections mainly focus on model development and broach relevant data-related considerations. Other stages are part of future work.

We outline different types of AI in section 2.1.1. Next, we emphasize the importance of data, domain embedding and an evolutionary view on growing[1] [2] intelligent systems in section 2.1.2. Section 2.1.3 sheds light on the deep learning algorithm, explaining how deep neural networks (DNN) learn and are trained, before briefly introducing AI challenges towards a responsible implementation in section 2.1.4. The following section 2.2 particularly highlights how to interpret AI output as a key challenge for AI lifecycle design. First, we outline metrics for model evaluation in section 2.2.1. Next, section 2.2.2 introduces the field explainable AI (XAI). Finally, in section 2.3.1, we highlight existing AI pitfalls focusing on model development. We emphasize the need for guidelines to enhance lifecycle design and introduce best practices, design patterns, and standards in the context of AI.

[1] This description of human interaction with AI was first heard during listening to an interview by Lex Friedman with Anthropic's team, and during that interview, AI researcher Chris Olah used this metaphor, which we liked.

[2] https://www.youtube.com/watch?v=ugvHCXCOmm4

2.1. Artificial Intelligence

AI is an umbrella term encompassing various computational approaches that enable machines to perform tasks typically requiring human intelligence. These tasks include perception (e.g. visual or audio-based), reasoning (e.g. large language models (LLM)), decision-making (e.g. clinical decision support (CDSS)), and continuous learning (e.g. the opportunity to implement on-site learning).

> In Article 3.1 of the European legislation on AI, the AI Act, an *AI System* is defined as "[...] a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" [Fut24].

Generally, AI system creation relies on structured processes, shared methodologies, and use-case-specific implementations. A successful, efficient and desirable realization of the AI lifecycle, from various view-points, including ethical considerations, is not trivial for the multitude of possible AI application scenarios. Therefore, equally with regard to regulation, shared common knowledge on RAI design decision-making along the AI lifecycle is an enabler of quality, which we should strive for.

This section introduces AI as a technology, aiming to highlight its complex inner workings under the inclusion of interdependencies with the real world, as well as internally from a more technical viewpoint, resulting in horizontal, i.e. AI in general, and vertical, under consideration of domain knowledge, views on intelligent systems.

2.1.1. Types of Intelligent Systems

Globally, AI systems can be categorized based on functionality and learning methodology, highlighting technical assets, as well as application to different use cases. This section provides a brief introduction how AI is used today. Throughout the present thesis, we focus on Deep Learning (DL) methods that enable human-like capabilities. A broad classification oriented towards the historical evolution of AI includes:

- *Rule-Based Systems* Logic-driven systems using predefined rules. For example, traditional, deterministic software is based on IF/THEN rules to produce output.

- *Machine Learning (ML)* Data-driven systems capable of pattern recognition. A popular non-DL-based example are decision trees [Qui86].
- *Deep Learning (DL)* Advanced deep neural networks (DNN) enabling hierarchical feature extraction. A sub-group of ML methods that enable human-like capabilities such as "listening" or "seeing". Popular examples comprise convolutional neural networks (CNN) [Fuk21], foundational to the field computer vision [SD20], which enables functionalities such as object detection, 3D reconstructions, as well as segmentation tasks, for instance.
- *Generative AI (GenAI)* Models that create new content, such as text or images, based on learned data distributions. GenAI comprises a subgroup of DL and is the main reason why AI became so popular as of toda. For instance, variants of transformer networks [VA17] currently count among the most popular architectures. For instance, Chat-GPT, a well-known LLM is based on the transformer architecture.

2.1.1.1. Historical Evolution of AI

"[W]hat is now known as artificial intelligence (AI) has evolved over more than two centuries in a long series of steps" [GA24b, 221]. The invention of relevant mathematical concepts reaches as far back as the 19th and early 20th century, famous names include Ada Lovelace, and Alan Turing [GA24b, 223], for instance. Throughout the 20th century, the field is characterized by a multitude of advancements, with artificial neural networks being "[...] first described by Warren Sturgis McCulloch, and Walter Pitts" [GA24b, 225] in 1943. In response, reflecting early considerations on AI, Isaac Asimov published his *Three Laws of Robotics* in 1950 [GA24b, 225]:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its existence as long as such protection does not conflict with the First or Second Law.

Shortly after, the first AI programs were introduced focusing on playing chess [GA24b, 225]. CNNs, influenced by well-known scientists Kunihiko Fukushima and Yann LeCun as well as popular concepts such as the ReLU (Rectifier Linear Unit) activation function for model architecture design were introduced during the second half of the 20th century. For instance, CNNs, whose architecture is based on "[...] neurophysiological findings on the visual system of mammals" [Fuk21, 1], are still optimized today [Fuk21, 1]. Finally, driven by computational advancements, the field experiences rapid growth since the 2010s, particularly in DL and neural network architectures. For instance, generative adversarial networks were introduced by Ian Goodfellow et al. in 2014. [GA24b, 226] A notable, slightly more recent development is the transformer architecture [VA17], introduced by Google Brain researchers in 2017, which has become foundational for GenAI.

As a popular example illustrating the rapid advancements of AI, in 2018, DeepMind's AlphaFold made substantial progress in predicting protein structures, a long-standing challenge in molecular biology [SA20]. By 2021, AlphaFold 2 [PJ21b] achieved accuracy levels comparable to experimental methods during protein discovery, leading to its recognition as a pivotal breakthrough in computational biology.[3]

Shortly after, in 2022, AI entered the consumer market and ChatGPT was released [GA24b, 226]. Since then, AI became very popular and experiences a hype. How this disruptive technology will shape our world is to be discovered, and a multitude of different related challenges emerge. For instance, its impact on consumers and user behavior [LW23] comprises a crucial research direction when aiming to evaluate the technology's impact on society, in addition to technical research further improving model outcomes towards RAI.

2.1.1.2. AI Application Today

As of now, AI is revolutionizing almost all industries, including agriculture, education, healthcare, finance, entertainment, transportation, military, manufacturing, gaming, food, drug discovery, e-commerce, and chemical industries [AM24b] [MS24]. Existing methods, if enough data is accessible, are basically applicable to any use case and domain, as well as perspective. For instance, AI can be applied to enhance technical aspects, such as the neural processing unit (NPU) [Lee21], that enables individual computers to run computing intensive DNN training (and inference). Another example comprises AI-powered bots that are transforming business operations, particularly in customer communication, information and knowledge management, and partial automation. They assist with tasks like handling Jira[4] tickets and collaborating with humans to streamline workflows. However, interoperability with existing systems, such as SAP [5], remains a major challenge, alongside finding sustainable pricing models for AI bots.

[3] As of now, they published AlphaFold 3, which is generalizable to other molecular structures, further highlighting the field's quick evolution.

[4] `https://www.atlassian.com/software/jira`

[5] `https://www.sap.com/germany/index.html?geotargering_redirect=true`

Overall, noteworthy improvements comprise promises of efficiency enhancements, cost reduction and improving operational performance. AI-driven solutions optimize existing processes, and may discover novel approaches to solving current challenges, such as e.g. lower energy consumption or the optimization of experimental approaches to medical problems, as is the case with AlphaFold [SA20] [PJ21b]. [MS24, 1] AI's impact and evolving role across sectors, emphasizes the need for a responsible integration with policy and industry practices [AM24b, 1]. Despite its benefits, AI bears multi-dimensional risks that need to be identified and mitigated. For instance, collaboration among experts is essential to ensure AI's sustainable and effective implementation across various use cases. This necessitates the establishment of interdisciplinary communication channels. Further, AI tends to deepen existing (not necessarily desirable) tendencies that are tied to ethical and societal questions. For instance, adoption remains concentrated in modern economies, underscoring the need for broader integration across industries worldwide. [MS24, 1] In summary, these advancements underscore AI's expanding role across various applications, driven by continuous research and innovative scenarios, starting with big data and a solid data strategy.

2.1.2. The Importance of Data, Domain Embedding & Continuous Conceptualization

The concrete implementation of individual intelligent systems varies based on the use case and domain-specific requirements, emphasizing the need to integrate use case-specific AI risk mitigation. Generally, intelligent systems share lifecycle processes, including data preprocessing, model training, evaluation, deployment, and maintenance-related design decisions. How they are realized impacts the overall intelligent system and its behavior. AI reliability in real world applications depends on a variety of different factors, including computing power, data quality, contextual understanding (highlighting AI risk evaluation), and ongoing adaptation to the system's surrounding (required based on the technology's evolutionary and stochastic character). Overall, key considerations when designing AI comprise:

- *Computing Power:* ML relies on computational power, as it operates on chips and semiconductors. Computing architectures determine the speed of training and inference, directly impacting the technology's advancement. Hardware not only influences speed but also shapes the design and development of ML models. Factors like chip power consumption determine where and how AI can be deployed in real world applications. [Hwa18, 1] Further, compared to other key AI inputs like data and algorithms, compute can serve as a particularly effective leverage point for regulation and governance. Governments and companies began to leverage it as a regulatory tool by investing in domestic compute capacity, restricting access to rival nations, and subsidizing sectors. [SG24a, 1]

- *Big Data & ML Preparation:* Big data plays a crucial role in enabling AI by providing the vast amounts of information needed for analysis, learning, and decision-making. As a result, big data and AI (re-)shape entire industries with the multitude of new applications they enable, equally in combination with other emerging technologies [DH23]. For instance, big data in healthcare is rapidly expanding due to the adoption of electronic health records (EHR) and standardized data exchange formats like DICOM and FHIR. However, challenges remain, including the need for standardized data labeling, improved data sharing, and open access to AI models via Application Programming Interfaces (API), for instance. [WS20] Overall, AI is a useful tool to interact with big data, aiming to optimize information extraction [BMS21], and it is crucial to approach the technology's foundation in big data in a structured manner when designing data sets. For instance, in [KY24], the authors propose to extend *the four Vs of Big Data*[6] with the "[...] six additional dimensions: value, validity, visualization, variability, volatility, and vulnerability" [KY24, 1]. Overall, these attributes show a strong pull towards the inclusion of domain knowledge, e.g. *value* highlights enhanced decision-making based on the generated output, which happens within a particular context, and *visualization* describes the transformation of outputs tailored to different stakeholders [KY24].

- *Data Quality:* Continuing considerations on how to design big data sets, implementing data quality requirements plays a crucial role. Data quality is multi-dimensional and is addressed from various angles [IEE22] [BG20] [RK23]. Considerations are centered around ensuring accuracy, consistency, and relevance of the data with respect to the intended use, underlying architecture and hardware within a dynamic real world. Additionally, privacy-related questions, and regulatory requirements, as introduced e.g. in the General Data Protection Regulation (GDPR) within the EU, comprise relevant sources of information for data quality.

- *Domain-Specific Embedding:* In the context of ML, the underlying data forms the "window" to the real world. The intelligent system generates output on new data samples it encounters through learnt patterns in the data, in addition to algorithmic design decisions carried out by contributing stakeholders along the AI lifecycle. Thus, its real world setting is impacted. Therefore, understanding all forms of relevant data, within their specific application context is crucial to ensure a responsible application. For instance, the data set plays a crucial role when designing the lifecycle in a way to enhance

[6] *The four Vs of Big Data* comprise *volume*, i.e. the amount of data, *velocity*, meaning the speed of data generation, *volatility*, the rate of data changes, and *veracity*, which addresses data quality and related trust questions [KY24].

model generalizability [MF22], anticipate data drift [RK23], or when interpreting performance evaluation metrics for a particular use case [HSA22], among other scenarios.

- *Continuous Monitoring:* Finally, since AI, focusing on DNNs, is non-deterministic in its behavior and opaque by nature, establishing mechanisms for continuously tracking model behavior is crucial to ensure a responsible innovation integration. A model monitoring strategy should be equipped to detect performance drift over time, possibly kick-start model updates, enable user feedback and mitigate AI risks, among other criteria. The EU AI Act identifies requirements for high-risk post-market monitoring in Article 72, highlighting strategy planning and documentation [Fut24].[7] The concrete and conform implementation is currently not 100% clear, and pre-determined change control plans [PT23], for instance outline possible approaches. As introduced in [ZV22], post-market monitoring specifically in healthcare should address "[...] collecting user feedback, technical monitoring and clinical validation" [ZV22, 1].

2.1.3. The Deep Learning Algorithm

After introducing relevant foundations for AI design, this section focuses on the heart of DNNs, i.e. the Deep Learning (DL) algorithm. DL serves as the backbone of modern AI models leveraging multi-layered artificial neural networks. At their core, DNNs are very complex, multi-dimensional, mathematical optimization problems based on stochastic weight optimization along all *neurons* that construct individual DNN *architectures*, which are loosely derived from our understanding of the cerebrum, or brain.

DNNs are at the center of our analysis, mainly thanks to the multitude of new horizons they enable. AI offers powerful data-driven insights in the age of Big Data, uncovering patterns and correlations that may be imperceptible to humans. It excels at processing complex, high-dimensional data, enabling multi-modal applications that are capable of executing human-like tasks. However, its effectiveness is often highly specific to the use case, as models rely heavily on the data and context they are trained on. Additionally, AI systems frequently operate as black boxes[8], lacking transparency in their decision-making processes. Since AI

[7] Further, the medical device regulation (MDR) outlines post-market surveillance system requirements for medical devices in Article 83 [Eur17].

[8] Fun fact: In DL, *black boxes* refer to systems whose internal decision-making processes are opaque or difficult to interpret. This is ironically the opposite of *black boxes* in aviation, which are specifically designed to provide maximum transparency and traceability after critical incidents. In a way, the aviation *black box* is the kind we do want in AI, one that reveals, rather than conceals, what happened and why.

thrives in data-rich environments, it is crucial to stay aware of where and how data is collected and the implications of its use. In addition, understanding the underlying mathematical concept that results in DL, provides a relevant piece in the repertoire of indirect methods that exist, aiming to shed light on the black box.

2.1.3.1. Foundation

The human brain consists of billions of neurons, the brain's primary functional unit. *Neurons* are specialized cells that transmit signals enabling sensations, movements, thoughts, memories, and emotions. These signals flow through complex, dynamic neural networks, forming the foundation of all brain activity. Neurons receive signals through tree-like extensions called *dendrites*. These signals are processed and transformed into electrical impulses, which are then transmitted via *axons* to other neurons through *synapses*, i.e. the key structures enabling communication in our nervous system. [Natb] [CM25] From a medical view, our current level of knowledge on the cerebrum is still in its infancy, with having experienced "major advances in our understanding of synaptic function" [SM08, 1] since the end of the 20th century [SM08]. Refer to [Süd18] for more information. Inspired by this biological process, DL models replicate neural connections mathematically and implement them technically. The combination of these artificial neural networks with large datasets is what powers modern AI.

The *perceptron* is a fundamental building block of "understanding" AI and serves to interpret how neurons in a DNN reflect human neurons. The basic *tunable input - calculation - tunable output sequence* is applied to all neurons that constitute a DNN, while differing mathematical functions for information processing exist. The perceptron takes a vector of real-valued inputs (the input signal for the dendrites), computes a linear combination of these inputs (signal processing based on mathematical computations instead of electrical signals), and applies a threshold function to determine its output (transmitted output to other neurons): 1 if the result exceeds the threshold and -1 otherwise. Next, for information passing to other neurons, tunable *weights*, real-valued constants, are applied to each output relating to an input sequence for other neurons, influencing "[...] the contribution of input xi to the perceptron output" [Mit97, 86]. Mathematically, the perceptron represents a hyperplane decision boundary in an n-dimensional space, dividing input instances into two categories. If the data points can be separated by such a hyperplane, they are considered linearly separable. However, not all datasets satisfy this condition. Perceptrons can model basic Boolean functions like AND, OR, NAND, and NOR, but they have limitations. For instance, a single perceptron cannot represent the XOR function, which requires a more complex, multi-layer approach. This limitation led to the development of multi-layer neural networks and more advanced learning algorithms. [Mit97, 86]

There is a substantial body of literature on this topic, describing foundations of DL with more profoundness. Tom Mitchell's classic book *Machine Learning*, published in 1997 [Mit97] is a foundational text that shapes modern AI and DL. It introduces core ML concepts such as inductive learning, decision trees, and neural networks, providing a rigorous yet accessible framework for understanding how machines "learn" from data. Many principles from this book, like hypothesis space, generalization, and bias, remain fundamental to today's DL advancements. The book bridges theory and practice, introducing the foundations of ML. Another popular book is *Deep Learning*, published in 2016 by Ian Goodfellow et al. [GI16]. It provides a comprehensive description of existing approaches to implement DL methods for various application scenarios from a conceptual view. In part I, they introduce mathematical foundations, which comprise relevant knowledge for adequately interacting with AI. Next, part II introduces basic neural networks and explains how DNNs learn and are trained, including more practical application scenarios. Finally, part III highlights the field's quick dynamics and introduces several DL-centered research directions, highlighting the currently very popular Deep Generative Models, for instance.

2.1.3.2. Key Concepts

"Machine learning [ML] is essentially a form of applied statistics with increased emphasis on the use of computers to statistically estimate complicated functions and a decreased emphasis on proving confidence intervals around these functions" [GI16, 96]. DNNs comprise a subgroup of ML methods (with their roots being the previously introduced perceptron) that are all centered around DL, which is most commonly based on backpropagation (optimization method to calculate the gradients) and gradient descent (optimization algorithm to adjust the weights) [GI16, 96], resulting in a very complex optimization problem, which we will explore in this section, introducing fundamental concepts.

Generally, two types of learning algorithms exist, i.e. *supervised* and *unsupervised* approaches[9] [GI16, 103] that differ in their algorithmic structure, including a different data setup. For the former, the data set consists of features and labels, while the latter does not provide labels and leaves it to the algorithm to discover patterns. Both methods are powerful, depending on the use case. For instance, anomaly detection in cybersecurity is an unsupervised optimization problem,

[9] One could argue that *Reinforcement Learning* (RL) comprises a third learning paradigm, since its approach, and terminology are fundamentally different. RL enables an autonomous agent to learn optimal actions by interacting with its environment, where the data is collected. The agent receives rewards or penalties based on its actions, guiding it towards goal achievement. This approach applies to diverse tasks like robot control, factory optimization, and game playing. The key challenge is learning from indirect and delayed rewards to maximize cumulative success over time. [Mit97, 367] These methods are not at the focus of our present contribution.

while COVID-19 segmentation in lung computer tomography (CT) scans needs to provide a data set including annotated lung segments.

The most popular optimization algorithm, *gradient descent*, is based on a gradient estimation method, commonly *backpropagation*, that calculates the gradients according to which gradually, the billions of *weights* in the neural network are optimized [GI16, 96]. In addition, enabling learning, a combination of *hyperparameters*, such as a *loss/cost function* and *learning rate*, serve to exercise influence on the model's bahavior in an indirect manner. The *training stage* is closely monitored and a collection of *epochs* are executed successively on a *validation data set*. After model training, global model qualities, such as its *generalizability* are evaluated on *test data*. A brief overview of DL-related core concepts, in addition to the data [GI16, 97], includes:

- *Network Architectures* Neural connectivity differences lead to specialized model architectures that are tailored to different tasks. Generally, DNNs, that consist of a combination of (different) layers of connected neurons, are designed to handle various problems by transforming complex, nonlinearly separable input data into more linearly separable features through multiple hierarchical layers. These layers combine linear and nonlinear functions, with neural *activation function*s (AF) playing a key role. They introduce non-linearity through transforming the input individual neurons receive. Common AFs include e.g. *Sigmoid*, *Tanh*, or *ReLU*, each contributing to the network's ability to learn and represent patterns effectively. [DS22b, 1] In addition to individual neuron's setup, the way how neurons are connected impacts performance. Popular architectural directions include:

  - *Convolutional Neural Networks (CNN)* [GI16, 330] are primarily used for image processing. They are characterized by a long history [GA24b, 225] and became really popular in 2012 with AlexNet [KA12] that displayed superior performance in image classification on ImageNet data [DJ09]. Simply put, they are based on a kernel that "folds" the input image into smaller dimensions and then unfolds them again to recognize features in the pixel/voxel data and assign them with meaning. They are introduced in more depth in Goodfellow et al. [GI16, 330].

  - *Recurrent Neural Networks (RNN)* [Sch19] *& Long Short-Term Memory (LSTM)* [GI16, 408] are a type of neural network designed to process sequential data (e.g., time series, speech, text). Unlike traditional networks, which treat each input independently, RNNs [Sch19] have a feedback loop that allows information from previous steps to influence the current step. This enables them to learn context and recognize patterns over time. LSTMs are a special type of RNN designed to enhance remembering information by introducing gates that control the flow of information [GI16, 408]. See Goodfellow et al. [GI16, 373].

- *Transformers* [VA17] are a powerful type of neural network architecture designed to handle sequential data without relying on recurrence (like RNNs and LSTMs). Instead, they use a mechanism called *self-attention*, which allows them to process all input tokens in parallel rather than sequentially. Thus, they allow for better long-range dependencies, and since their publication in 2017 [VA17], they became state-of-the-art models for natural language processing (NLP) [DJ19] and other generative tasks.

- *nnUNet [IF21] & Segmentation Variants* A popular segmentation model variant is nnU-Net (no-new-Net), an advanced DL framework designed for automatic medical image segmentation. Unlike traditional U-Net models, nnU-Net [IF21] does not require manual tuning. Instead, it automatically configures itself, e.g. pre- and post-processing, based on the dataset it is given. In addition, the framework accepts various input data formats and surpasses manually designed segmentation networks. Segmentation models, and especially, the nnUnet comprise the foundation for our evaluation in chapter 8, which is embedded within a medical software, as introduced in section 5.2.2.

- *Backpropagation & Gradient Descent* comprise the fundamental mechanisms for training DNNs through forward and backward calculations on the input features. The aim is to fine-tune adjustable weights along connecting neurons within the DNN architecture to minimize the calculated *loss* of the *loss function*. Therefore, the *optimizer* Gradient Descent, or e.g. its popular variant Stochastic Gradient Descent (SGD), calculates *gradients*, i.e. directions on how to adjust *weights* towards optimizing the model's output within the *search space*. Through various iterations, i.e. *epochs*, the *learning rate* determines the algorithm's "step size", and other hyperparameters, such as e.g. *momentum* monitor characteristics such as its "speed". Refer to [GI16, 80] [Mit97, 89], for more information. The algorithm is capable of determining local minima through analyzing different *batches* of the complete training data set with adjustable size. For most practical use cases, this approximation on the training data suffices. However, *weight initialization* influences the output, and techniques such as *transfer learning* [FZ19] emerged to provide more targeted methods than random initialization. The latter comprises only an example illustrating the complexity of DNN architecture design. In general, a multitude of different (dependent) *hyperparameters* can be adjusted to individual application scenarios, resulting in a combination of design decisions.

- *Hyperparameters* are configurable settings that influence how a DNN model learns. Unlike model parameters, or *weights* (which are learned from data), hyperparameters must be set before training begins [GI16, 96], possibly including an updating strategy aligned with the training progress. Choosing the right hyperparameters is crucial for optimizing model performance and preventing *overfitting* (reflects the training data distribution and cannot generalize) or *underfitting* (not capable of discovering patterns) in the data. The following list highlights three relevant key hyperparameters for DL, while many more adjustable configurations exist, such as e.g. the previously introduced *epochs* and *batch size* equally focusing on model training, or the *activation function* within a DNN architecture:

- *Loss/Cost Functions* measure how well a model's predictions match the actual values or goal. It provides feedback to update the model's parameters during training. The aim is to minimize the loss, improving performance. For instance, during classification tasks, a popular example comprises the *cross-entropy loss* (CE) [MA23], which measures the difference between predicted probability distributions and true labels. Popular variants of CE, such as e.g. the *focal loss* [Lin18], originating from object detection with a high background pixel count, adjust the loss function so that it is more suitable for imbalanced data sets through weighing majority class-related loss less and punishing the loss calculated on rare samples, or pixels comprising the object to detect, stronger. The loss function plays a central role during model training and careful considerations, in alignment with domain knowledge are required for design decision-making. For an overview of common possible choices with respect to different tasks, refer to [TJ24].

- *Optimizers* update a model's parameters (weights) to minimize loss and improve performance. They determine how the model learns from data by adjusting weights based on gradients, i.e. "[...] weight vectors to find the weights that best fit the training examples" [Mit97, 89]. A popular method is Gradient Descent [Rud17], which "[...] provides the basis for the BACKPROPAGATION algorithm, which can learn networks with many interconnected units" [Mit97, 89]. Its popular variant SGD updates model weights incrementally using small batches of data, instead of computing gradients over the entire dataset [AI21]. Another common variant is the Adaptive Moment Estimation (Adam) optimizer [KB17]. Adam combines the advantages of SGD with momentum and the optimization method RMSProp [KB17, 5], adapting learning rates for each parameter based on past gradients.

- *Learning Rate* is a crucial hyperparameter in ML that determines how much a model updates its weights during training. It controls the step size taken in gradient descent to minimize the loss function. If it is too high, the model might overshoot the optimal solution in the search space, leading to instability. In case it is too low, the training can be

> too slow or get stuck in some local minima. "It is usually set to some small value (e.g., 0.1) and is sometimes made to decay as the number of weight-tuning iterations increases" [Mit97, 88].

This list comprises only an excerpt of relevant design decisions surrounding the setup of AI for the multitude of possible scenarios. Other interesting concepts to consider are e.g. centered around *regularization* [FS22], a technique used to prevent overfitting and ensure models generalize well to unseen data. Or, other (combinable) techniques like *dropout* [SN14] randomly deactivate neurons to prevent over-reliance on specific features. Generally, *hyperparameter-tuning* and overall model optimization ensure models are more robust, generalizable, and reliable for real world applications.

Our aim was to introduce the core idea of DL in this section, highlighting emerging intricacies of relevant design decision-making. For more precise information, refer to existing literature [GI16] [Mit97]. The next section explores dependencies between solving these complex optimization problems at the core of DL and currently available compute.

2.1.3.3. Mathematics & Computation

The previously introduced optimization problem and stochastic learning, fundamental to DNNs result in challenges regarding currently available computational power and the identification of the global optimum for individual optimization problems. In [BJ22], the authors introduce *The Modern Mathematics of Deep Learning*. This mathematical analysis of DL explores fundamental questions beyond classical learning theory. Key topics include the exceptional generalization of overparameterized networks, or the role of depth in DNNs, among other questions. This field develops modern approaches to address identified challenges rooted in the conceptual setup and technical implementation of DNNs. AI models operate stochastically and e.g. many classification problems are not Turing computable [BH23] [LY23], which describes the state when a machine is capable to solve any computation problem.

Today, almost all calculations are carried out on digital machines, which strongly impacts performance. "The question whether a continuous system can be simulated on digital computers lies therefore at the heart of the foundations of signal processing and computer science" [BH23, 1]. Novel approaches, such as quantum computing will enable more powerful calculations, which will impact future AI output.

2.1.4. Challenges in AI & Responsible Implementation

While AI presents vast opportunities, its creation, integration and application introduce significant challenges centered around legal, ethical and technological risks [KA21]. In addition to the previously introduced methodological complexities, challenges are related to the intelligent system's intended real world setting, and depend on the *intended use*, as well as possible *misuse* scenarios, considering risk management (RM) [ISO19]. Particularly, in high-risk domains, such as medicine, methodological missteps when designing AI lifecycles may result in grave consequences if undiscovered. Healthcare-specific challenges are outlined in more detail in section 5.1. Generally, a responsible implementation should be designed carefully, with close ties to the use case and addressing AI's evolutionary character to achieve the intended results. Additionally, this new technology's impact on societies is not yet fully clear, resulting in complex ethical questions when integrating AI with the real world. Responsible AI (RAI) is detailed in chapter 3, where we introduce the vision of implementing AI, which is ethical, lawful, and accountable [DRN23], highlighting the central role of implementing an AI risk management system (RMS) [Org23], which is defined in Article 9 of the AI Act [Fut24], the European regulation on AI.

In summary, this section aims to provide a comprehensive foundation for understanding AI from a methodological viewpoint, highlighting its intricate inner workings, as well as resulting technical challenges for RAI implementation and real world integration. The next section 2.2 focuses on interpreting AI output, a challenging endeavor regarding the technology's inherent opacity within a dynamic real world. Namely, we introduce AI performance evaluation and Explainable AI (XAI) in more detail, with respect to our use cases in part III.

2.2. Examining AI Output

Evaluating AI models requires not only task-specific performance evaluation metrics [ISO22b] [MD22] [SE10] [RA21], but also methods to evaluate further, conceptually more complex quality criteria, such as usability [Bro95], fairness [ISO21a], transparency, or reliability [HA22], among others. Additionally, purely statistical analysis of the model's architecture and its components contributes relevant information, e.g. addressing situations where access to the model's development setup is not possible [MC21]. We put a particular emphasis on classification performance evaluation metrics with respect to our multi-label ECG (ECG) classification use case in section 5.2.1 for our evaluation in chapter 7, which introduces metrics in more detail.

Generally, the true success of any application can only be measured in the complex and dynamic real world. These intricacies have led to the development of numerous approaches designed to improve overall AI system transparency, aiming to facilitate a responsible integration into real world applications. Explainable AI (XAI) provides a toolbox on how to shed light on the model's inner workings from different angles. Various publications characterize this quickly evolving field [LH16] [BG20] [DS21] [HA23] [Leb23] [BX23] [BC24]. In the present section, we introduce XAI with a particular emphasis on SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) explanations, which comprise the foundation for our evaluation of MQG4AI in chapter 6.

In summary, beyond numerical evaluation, understanding and interpreting AI outputs is crucial for ensuring reliability and informed decision-making within the respective real world workflow. This section summarizes key considerations in AI output interpretation, emphasizing the balance between quantitative evaluation in form of (performance) evaluation metrics and qualitative aspects like explainability.

2.2.1. AI Performance Evaluation Metrics

True success in AI evaluation can only be measured in the real world, a nontrivial challenge that depends on both, the structural setting (e.g., GenAI, classification, or segmentation models) and domain-specific knowledge, which indirectly impacts evaluation interpretability. Choosing the right performance evaluation metrics is a crucial design decision, as they provide fundamental information for further decision-making along the AI lifecycle. Additionally, their correct interpretation depends on the respective recipient's knowledge among involved stakeholder roles. Incorrect choices can introduce significant risks and pitfalls [HSA22] [TJ21]. Therefore, it is important to invest research time in metrics selection for a reliable evaluation within the intended real world setting, e.g. refer to [RA21] for a thorough analysis of popular metrics for image-level classification, semantic segmentation, instance segmentation or object detection. Or, in [MD22], the authors provide a thorough analysis of popular metrics for medical image segmentation. These examples highlight a possible approach for metrics analysis, defining the starting point for metrics selection based on project-specific criteria. This section first outlines general considerations surrounding reliable metrics, before briefly introducing classification performance evaluation metrics in more detail.

2.2.1.1. Towards Reliable Metrics

Ensuring a reliable (performance) evaluation of the intelligent system requires careful consideration of relevant information infrastructures. For instance, in addition to data and domain-related considerations, key aspects include effective communication channels between disciplines within a team that consists of domain experts, developers, and other stakeholders. This setup results in the need to address the critical question *Who needs to know what, and when?*, tailored to specific design decisions such as (performance) metrics selection. In chapter 7, we provide a comprehensive analysis of (multi-label & binary) classification metrics-related risks and relevant research directions that extend the following, more general considerations, which comprise the foundation of reliable (performance) evaluation.

Standards serve as a foundation for consistency, guiding both evaluation methodologies and reporting practices. For instance, certain classification metrics are asymmetric, meaning that the designation of a class as positive (1) or negative (0) influences their outcome and cannot be arbitrarily switched [HSA22, 3]. To ensure consistency in binary classification within healthcare, it is reasonable to adopt a standardized convention where the disease is defined as the positive class and healthy samples as the negative class, as proposed in [TJ21, 9]. Moreover, inconsistencies in metric selection should be addressed by establishing a standardized collection of evaluation metrics for auditing various ML applications like (medical) image classification or segmentation. This would complement "[p]eer-reviewed randomised controlled trials as an evidence gold standard" [KC19, 2], rigorously assessing potential risks and clinical efficacy. The European standardization landscape surrounding AI is detailed in section 3.1.3.2.

Benchmarking Quantitative AI benchmarks are essential for assessing model performance, capabilities, and safety, increasingly shaping AI development and regulatory frameworks [EM25d, 1]. Benchmarking trained models across various (medical) domains is required to establish a widely accepted evaluation baseline. For official benchmarking, one approach involves creating independent real world test sets that remain publicly unavailable [KC19, 3]. Alternatively, simulation studies using synthetic data offer another viable strategy [FF24, 3]. Furthermore, the development of platforms that provide the necessary infrastructure is crucial to facilitate standardized benchmarking practices, including use case-adapted metrics' compilations for evaluation. However, the growing influence of benchmarking strategies raises concerns about their effectiveness in evaluating high-impact capabilities, safety, and systemic risks. Issues range from biases in dataset creation and inadequate documentation to systemic problems such as misaligned incentives. Additionally, many benchmarks fail to account for AI's multi-modal nature and real world interactions. [EM25d, 1] Also, careful consideration is required regarding metrics selection, as they exhibit varying behavior depending on the data collection process and the resulting diversity [TJ21, 5]. Beyond technical flaws, benchmarking is influenced by cultural, commercial, and competitive dynamics, often prioritizing state-of-the-art performance over broader societal concerns [EM25d, 1]. In summary, "[...] not all benchmarks are the same: their quality depends on their design and usability" [RA24, 1], which needs to be established in alignment with different (groups of) use cases.

Clarity of Measurement A majority of quality criteria for RAI, such as *Transparency* or *Robustness* and *Performance*, are grounded in Trustworthy AI (TAI), which is introduced in detail in chapter 3. Challengingly, the interpretation of these rather abstract and complex concepts depends on a multitude of different, use case-specific factors. They are not trivial to accurately measure with respect to individual applications. For instance, our interpretation of *fairness* and the multi-faceted concept of *bias* are strongly influenced by a variety of different and fluid factors, which is why we introduce *Transversality* in [EM25b], as further outlined in section 3.2.2.3. While fairness metrics exist [MM24b], they might not be suitable for every use case and their applicability needs to be evaluated. Generally, it is crucial to define what is being measured and to identify or develop metrics that accurately reflect these criteria within the respective context. "The key concept is that the metrics are chosen so that actions and decisions which move the metrics in the desired direction also move [...] desired outcomes [in the real world] in the same direction" [HK98, 6]. For instance, in [HK98], which was published in 1998, the authors introduce seven steps to identify good metrics, highlighting domain embedding and the identification of interdependencies, which is still relevant today. In case existing metrics do not suffice to adequately capture all desirable qualities, novel metrics are developed, which is often the case in biomedicine [TJ21, 9], as with the Physionet scoring metric [RM21a] for ECG classification or the S_{CASP} Score [PJ21a] to evaluate generated protein structures, including the previously introduced AlphaFold2 model [PJ21b]. Finally, the *Foundation Model Transparency Index* (FMTI) [BR23] serves as an example where the name is transferring a misleading message on what the index[10] actually

[10] Indexes are loosely understood as aggregated metrics: "[...] an index for a specific entity is the aggregate of multiple low-level indicators that can be more directly quantified" [BR23, 13].

does. *Transparency*, as discussed in sections 3.1, 3.2.2.2 and 3.3.5.4 is complex to grasp, encompassing concepts like *Explainability*, *Interpretability* and *Traceability*, as well as considering all elements that comprise the AI system [Eur19, 18]. The FMTI assesses "[...] transparency of foundation models with a comprehensive ecosystem-level approach [...]" [BR23, 13], primarily evaluating "[...] foundation model developers for their transparency [...]" [BR23, 24]. However, this focus on developer transparency means it does not fully address the (scientific) challenges of making opaque AI models transparent. The index "[...] focus[es] on structural forms of transparency, taking a more macroscopic perspective" [BR23, 13]. As possibly reliance on the index grows, other critical *Transparency* criteria risk being overlooked, while a system is believed to be transparent. Given these limitations, concerns about its potentially misleading title appear justified, and a possible enhancement results in *Foundation Model Developer Transparency Index*. As of 2024, developer transparency remains modest, with an average score of 58 out of 100, "[...] a 21 point improvement over FMTI 2023" [BR25, 1]. This still limited level highlights the need for structured approaches such as AI lifecycle planning templates to support more accountable and traceable AI development.

2.2.1.2. AI Classification Evaluation

In general, a compilation of metrics is necessary for a comprehensive overview of the model's performance, since "[n]o single metric captures all the desirable properties of a model [...]" [HSA22, 1]. This section briefly introduces classification performance evaluation, which is elaborated in more detail in chapter 7. For instance, ISO/IEC TS 4213 on the *Assessment of ML classification performance* [ISO22b] provides a comprehensive overview of existing metrics from a horizontal, technical viewpoint, highlighting a correct application/interpretation. Regarding classification metrics as a whole, two fundamental perspectives are established for evaluating model performance, depending on the available artifacts.

Confusion-Matrix Following this approach, the model's performance across different classes is evaluated using the *confusion matrix* [TJ21, 5]. This method relies on *thresholding* to classify predictions as either *true* or *false*. Its applicability, including widely used metrics such as *Accuracy*[11], *Recall*, *Specificity*, *Precision*, and *F1-score*, as well as common pitfalls in performance evaluation within current research have been extensively analyzed in [HSA22], for instance. Refer to section 7.2.4.1, where we detail an exemplary metrics compilation for ECG multi-label classification.

Ranking The ranking-based perspective relies on the real-valued function

[11] Accuracy is an inadequate measure for imbalanced datasets and should be replaced by *Balanced Accuracy* [TJ21, 5].

learned by the model, which returns confidence scores, thereby requiring access to the model. Moreover, this approach provides a threshold-independent performance evaluation, encompassing metrics such as the popular but highly discussed *Receiver Operating Characteristic Area under the Curve* (ROC AUC) [ZZ14, 1822], and can also be utilized for threshold optimization towards enhancing confidence. For a comprehensive discussion of ROC AUC's suitability on imbalanced data sets, refer to section 7.1.1, where we introduce metrics-related risks.

After shedding light on the evaluation landscape surrounding AI and guiding directions for consideration, with a particular emphasis on classification scenarios, the next section explores explainable AI, highlighting inherent intricacies of this rather new field.

2.2.2. Explaining AI (XAI)

Explainability is a key requirement for trusting AI outputs, addressing the inherent opacity of DNNs. Without methods to shed light on intelligent system's inner workings, "[...] we risk to create and use decision systems that we do not really understand" [GR18, 2]. The necessity of explaining ML models gained significant recognition within the AI community after 2017 [BA19, 3], complementing the evolution of AI into the powerful technology as it is commonly known today. Explainable AI (XAI) "[...] is foundational to the development of RAI systems" [BX23, 26], and plays a crucial role in developing trustworthy AI (TAI) [MA21b]. Existing XAI methods and tools [BS23] can be integrated into the broader framework of RAI regarding aspects such as fairness, robustness, privacy, and security [BX23]. For instance, *Explainability* is referenced from multiple perspectives in the "Ethics Guidelines for Trustworthy AI", as introduced by a High-Level Expert Group established by the European Commission (HLEG) [Eur19], which we outline in section 3.1.2. First, *Explicability* comprises an ethical principle contributing to the foundations of TAI [Eur19, 8], while TAI criteria such as *Transparency* and *Human agency and oversight* are closely linked to *Explainability*, as further outlined in sections 3.2.2 3.3.5 on TAI-related RM.

As a result of the field's novelty and complexity, numerous challenges remain, with key research questions including the establishment of generalizable guidelines for designing "good" explanations and the development of universal frameworks for assessing explanation quality. The vast range of potential application domains further complicates the generalizability of explanations, as effective interpretability necessitates tailoring to specific contexts. A crucial next step involves the practical integration of AI systems and their explanations into real world applications, while their tangible benefits have yet to be empirically validated [MA21b, 3].

This section first introduces relevant terminology to grasp the concept of *explainability* in section 2.2.2.1. Section 2.2.2.2 sheds light on explanation purpose and the relevance of domain embedding, as well as respecting the target audience. Next, section 2.2.2.3 introduces how to approach methods to implement XAI, introducing two popular explainability methods (LIME and SHAP) in more detail. The process of XAI method selection is outlined in section 2.2.2.4. Finally, section 2.2.2.5 presents the XAI method evaluation landscape in more detail, highlighting the complexity surrounding relevant explanation quality criteria. This section comprises the foundation for chapter 6, where we attempt to organize XAI-related processes along the AI lifecycle development stage in a responsible manner, including a technical method to evaluate explanation quality [LGC24].

2.2.2.1. Terminology

First, we introduce important concepts surrounding XAI. "Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand" [BA19, 6]. This definition highlights the complexity of defining a "good" explanation, which comprises a multitude of rather abstract concepts that are not trivial to implement. Challengingly, current research lacks a standardized terminology and instead relies on ambiguous and inconsistent definitions [MA21b, 3]. For instance, stakeholder interpretability plays a crucial role along the AI lifecycle, and system users require a different explanation setup than the application of XAI methods by developers to evaluate complex criteria such as model robustness [DA21] or data quality [BG20], for instance. Therefore, a crucial initial distinction must be drawn between *explainability* and *interpretability*, as these terms are often used interchangeably despite representing distinct concepts [BA19, 4]. The following list comprises a summary of XAI-related key concepts:

- *Interpretability* pertains to "[...] the ability to explain or to provide the meaning in understandable terms to a human" [BA19, 5]. It is defined as "[...] the mapping of an abstract concept (e.g., a predicted class) into a domain that the human can make sense of" [MG17, 2]. This can be regarded as a "passive characteristic" [BA19, 4], evaluating to which extent "[...] a given model makes sense for a human observer" [BA19, 4]. In other words, "[a] model can be explained, but the interpretability of the model is something that comes from the design of the model itself" [BA19, 6].

- *Explainability* extends beyond interpretability by encompassing "[...] any action or procedure taken by a model with the intent of clarifying or detailing its internal functions" [BA19, 5]. As such, it can be regarded as an "active characteristic" [BA19, 5]. This implies that supplementary information must be provided alongside the model to enhance its accessibility for various stakeholders. This requirement is also embedded in the definition of an explanation, which is specified as "[...] the collection of features of the interpretable domain that have contributed for a given example to produce a decision (e.g., classification or regression)" [MG17, 2].

- *Transparency* If a model can be understood by humans without requiring additional information, it is considered *transparent* [BA19, 5]. For instance, decision trees are inherently transparent, whereas DNNs function as *black boxes*, characterized by their *opacity*, as introduced in the previous sections. As a result, implementing model transparency for DNNs is a non-trivial task and comprises multiple indirect methods, such as XAI methods, or the documentation of AI lifecycle design decisions, for instance. Transparency in the context of RAI is discussed in more detail in chapter 3.

- *Understandability* "[...] denotes the characteristic of a model to make a human understand its function" [BA19, 5], and assesses whether a model's functionality is comprehensible to any human stakeholder without necessitating an explanation of its internal processes [BA19, 5]. It is summarized as "[t]he degree to which a human can understand a decision made by a model" [BA19, 5].

- *Comprehensibility* pertains to how an explanation is presented and whether it is in a form that humans can understand. It is defined as "[t]he quality of the language used by a method for explainability" [VL20, 9]. Additionally, for explanations to be useful, stakeholders must be able to grasp the model's functionality [BA19, 5], which requires them to either already have or be provided with the necessary (domain) knowledge.

This selection of relevant concepts highlights fine-grained characteristics that need to be considered for explanation design and evaluation, bridging model output and human interpretability. Generally, in the context of software development the recipient equally plays a crucial role, and valuable insights can be drawn from this domain. For example, the authors of [ARD09] introduce *Knowability*, i.e. "[...] the property by means of which the user can understand, learn, and remember how to use the system" [ARD09, 5], highlighting user interpretability in more detail in the context of *Usability*[12]. *Knowability* is further subdivided into a compilation of related concepts such as *Clarity*, *Consistency*, and *Helpfulness*, for instance, resulting in a comprehensive taxonomy describing the concept more thoroughly [ARD09, 5], which could equally serve as a foundation for designing and evaluating effective explanations.

[12] *Usability* is referred to in more detail in the next section when introducing the explanation evaluation stage, since it is identified as an important criteria for explanations, as well.

2.2.2.2. Purpose & Audience

"Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand" [BA19, 6]. Therefore, important considerations need to be made prior to explanation design and method selection, including stakeholder analysis and domain embedding. In addition to clarifying the explanation's purpose, relevant considerations focus on who applies which type(s) of explanation(s) at what lifecycle stage [DS21]. For instance, in [HS20], the authors propose to design explanations with respect to three criteria (user *Roles*, lifecycle *Stages*, and individual *Goals*) towards a successful explanation application [HS20, 7]. The following list summarizes important considerations that impact the XAI method:

- *AI Lifecycle* XAI methods can be applied at different stages along the AI lifecycle beyond human user interpretability [HS20] during model application in the real world. For instance, data-related design decisions include the evaluation of data quality [BG20] or optimized feature selection during preprocessing [ZJ22]. Generally, shedding light on the model's inner workings supports involved stakeholders during model development.

- *Objective* Overall, XAI aims to enhance the transparency of AI systems by making their decision-making processes and outcomes interpretable to human stakeholders. This is essential for ensuring TAI, which relies on key criteria that can be implemented using XAI methods. Examples comprise fairness [Leb23], robustness [DA21], and managing trade-offs, including the explainability-privacy tradeoff [SS24] and the XAI-accuracy tradeoff [van21]. Additional considerations aim to distinguish black box explanations from a more functional view. For instance, in [GR18], the authors define three categories: black box model explanation (1), where a simplified model mimics the black box's behavior [GR18, 13], black box outcome explanation (2), which focuses on making the model's predictions more understandable [GR18, 14] and black box inspection (3), which provides textual or visual insights into the model's inner workings or outputs.

- *Domain Embedding* XAI methods necessitate the inclusion of complex (medical) domain knowledge to result in "good" explanations. For example, in skin cancer detection, Stolz's ABCD Rule of Dermatoscopy [SH07] is used to explain the classification of lesions as melanoma (cancerous) or nevus (healthy). The XAI approach, as presented in [SF21], attempts to incorporate that domain knowledge to explain the model's predictions to the physician in charge through perturbed images, with the perturbations aiming to capture characteristics of Stolz's ABCD Rule. It involves comparing the model's predictions on original and perturbed images, with positive perturbations highlighting melanocytic features and negative perturbations simplifying them to resemble nevus features. Other domain-specific considerations for XAI in healthcare can be found in [BS23] [MA21b], and in [AZB23] for ECG use cases in particular, for instance.

- *Stakeholder* Depending on the explanation's function within the AI lifecycle [DS21], other explanation recipients than the human user exist. For example, developers are key stakeholders who interact with explanations and interpret their meaning. However, thanks to their more extensive background knowledge compared to AI system users, they require fewer additional explanation layers, thereby reducing the risk of misinterpretation. In summary, the prior knowledge of explanation recipients shapes the explanation design. Further guiding considerations when designing explanations can include additional, audience-related concepts. For instance, for the user in particular, aiming towards human-machine teaming [Hen22] can be beneficial for the real world integration of the intelligent system.

2.2.2.3. Methods to Explain Opaque Models

The present research focuses on black box models, or DNNs, which are models whose internal mechanisms are not directly understandable by humans without the use of additional explanation methods [GR18, 1]. A multitude of different XAI methods exist, and use case-specific considerations, as previously introduced, impact the design decision-making process, in addition to technical characteristics of method applicability, which we highlight in the following. The book *Interpretable ML* by Christoph Molnar provides a comprehensive introduction [Mol25] to XAI and existing methods that cover interpretable models, local, and global model-agnostic methods, as well as the interpretation of neural networks. The following list highlights key criteria to classify XAI methods for method selection:

- *Stage* XAI methods can be categorized based on their application stage within the model implementation process. **Ante-hoc** methods focus on designing inherently explainable models, while **post-hoc** methods provide explanations for already trained models. Post-hoc approaches are further divided into **model-agnostic** methods, which work independently of the model's structure, and **model-specific** methods, which are tailored to a particular model type. **Model-agnostic** methods are independent of the structure of the model to be explained, and thus can be combined with any type of black box model. In contrast, **model-specific** methods are tailored to a specific type of model. [VL20, 15] The majority of XAI methods focuses on *supervised learning*, while *un-*, *semi-supervised*, and *reinforcement learning* techniques equally necessitate explanations. [SF19, 2]

- *Scope* refers to the range of what the explanation explains. **Global** explanations aim to encompass the overall model functionality, and decision-making process. In contrast, **local** explanations generate an explanation per data instance at a time. [VL20, 15] In other words, the latter "[...] implies knowing the reasons for a specific decision [...]" [ZJ21b, 9], while the former "[...] means explaining why patterns are present in general [...]" [ZJ21b, 9].

- *Explanation Input* Post-hoc XAI methods expect model output as input, and the model's functionality shapes the structuring of explainability methods [VL20, 15], which in turn is impacted by the input data set, including data preparation steps.

- *Explanation Output* The selection of an appropriate XAI method is impacted by considerations on the desired output format [VL20, 15]. Different types include numeric, rule-based, visual and mixed explanations [VL20, 17], and the desired format is closely related to the respective recipients.

Model-specific methods comprise ante-hoc methods by design. They encompass various implementation strategies designed to inherently align the model with its intended domain towards explainability by design. Examples for these approaches include rule-based methods, such as those emphasizing neuro-symbolic AI [SZ21], as well as prototype-based layers [LO17]. A popular example for model-specific, post-hoc methods are Class Activation Mapping (CAM) techniques, that use saliency masks (SM) to identify the features most influential in a DNN's prediction [GR18, 29]. As a visual explainability method, saliency maps highlight the regions of an input image that contribute most to a given classification [VL20, 20]. A saliency mask is defined as "[...] a subset of the original record which is mainly responsible for the prediction" [GR18, 29]. However, the mostly model-specific nature of CAM methods limits their generalizability and requires new implementations for different problems [GR18, 29], as in [GS22] for explaining an ECG classifier.

In contrast, model-agnostic methods are extendable to multiple domains and application scenarios, since they do not depend on a particular model, which makes them generalizable by definition [GR18, 28]. In the following, we outline two post-hoc and model-agnostic methods in more detail. They are at the center of the example XAI method evaluation procedure in chapter 6.

Local Interpretable Model-agnostic Explanations (LIME), as introduced by [Rib16] in 2016, has gained significant popularity within the XAI community [GR18] [Gil18] [MA21b] [BP20] [VL20] [BA19]. The method's popularity may stem from its model-agnostic nature and independence from input data types, allowing it to "[...] return an understandable explanation for the prediction obtained by any black box" [GR18, 30]. LIME is based on a "locally interpretable model" [GR18, 29] to explain individual data samples. The algorithm approximates complex models locally using simpler linear models, as achieving a global approximation is often infeasible [Rib16, 4]. By exploring the local region around an instance, LIME generates explanations that reflect the model's behavior in that vicinity [Rib16, 3]. The method is perturbation-based, meaning it estimates the impact of features by modifying similar instances [RSB18, 7]. For image classification, LIME segments an image into "super-pixels" [Rib16, 4] and evaluates their importance in the model's prediction, ultimately highlighting only the relevant parts while greying out the rest. However, LIME does not account for feature interactions, and its explanations may lack faithfulness and stability [RSB18, 9].

SHapley Additive exPlanations (SHAP) is a numerical approach that quantifies the relevance of input features for model predictions [VL20, 17]. It builds on the idea that any explanation of a black-box model is itself an "explanation model", serving as an interpretable approximation of the original model [LSI17, 2]. Like LIME, SHAP belongs to the class of *additive feature attribution methods*, that "[...] link the input features with the model prediction, providing an interpretation based on a linear formulation of the models" [CA24, 1]. SHAP in particular is derived from cooperative game theory principles [LSI17, 3]. A key strength of SHAP is that it provides a unique solution satisfying three desirable properties [LSI17]:

- *Local accuracy* means that the input for the explanation model, and its output, correspond to the original model's prediction for a given input.
- *Missingness* implies that features missing from the input of the explanation model do not impact the correct prediction.
- *Consistency* states that the impact of an input feature should not decrease if the original model changes and that input feature's significance increases or remains unchanged, independent of other attributes.

Among additive feature attribution methods, only SHAP ensures all three properties [LSI17, 2], making it a unified measure of feature importance [LSI17, 4]. However, SHAP's main drawback is its computational cost, as it scales exponentially with the number of features, making it significantly slower than LIME for complex models [BH21, 15].

2.2.2.4. Method Selection

After outlining important considerations for XAI method selection, this section highlights the complexities of this iterative design decision-making process, which is in principle transferrable to other design decisions surrounding model development. In [BP20], the authors provide a concrete example of how a data scientist, Jane, navigates the complex decision-making process of selecting an appropriate XAI method for a credit eligibility ranking task. The goal is to optimize both model performance and explainability, balancing the trade-off between opaque models (which often perform better) and transparent models (which are more interpretable) [BP20, 17].

Model Selection Initially, Jane faces the decision of whether to choose a transparent model with high interpretability or an opaque one that may yield better performance and generalizability. After testing various models, she opts for a *Random Forest classifier*, an opaque model that optimizes overall performance [BP20, 17]. For this step to be executed, relevant interdependencies, emphasizing the input data and use case are first defined in more detail.

XAI Method Selection The process of selecting a XAI method involves first choosing the model and then determining which XAI technique(s) will best explain the model's decisions. Depending on the level of explainability needed and the target audience, Jane considers multiple XAI methods. She begins with SHAP to explain the model's decision-making process through feature importance. However, SHAP only provides local explanations for individual instances, which makes it challenging to assess the model's global behavior [BP20, 18]. Local explanations can differ significantly for different instances, and evaluating the clarity of local model-based explanations is challenging [MA21b, 15]. Clarity means that "the explanation is unambiguous, i.e. it provides a single rationale that is similar for similar instances [...]" [MA21b, 3]. As a result, Jane explores additional methods. This iterative process of applying various XAI methods, analyzing explanations, and posing new research questions continues until Jane has employed multiple XAI approaches [BP20, 19].

This example highlights that the process of selecting and combining XAI methods is dynamic, involving the enhancement of model performance, stakeholder involvement, and building trust with end-users [BP20, 27]. Moreover, the lack of clear guidelines on how to combine different XAI methods adds complexity to the creation of a cohesive XAI system [BP20, 20]. The authors outline five evaluation criteria, i.e. comprehensibility, fidelity, accuracy, scalability, and generality to guide this decision-making process [BP20, 4]. In the next section, we discuss explanation quality criteria in more detail.

2.2.2.5. Quality & Design

The ultimate goal of evaluating the quality of explanations is to determine "[...] to what extent the properties of explainability [...] are satisfied" [ZJ21b, 8]. However, defining a *good* explanation remains challenging, as the question "what is a good explanation?" is inherently subjective, depending on the specific context and its users [ZJ21b, 11]. The diverse objectives of various XAI approaches complicate the selection of suitable evaluation metrics. Given this complexity, it is often up to the development team to determine meaningful evaluation criteria and validate them in collaboration with relevant stakeholders [Gil18, 9]. Overall, the effectiveness of an explanation depends on how well it performs within its application domain, making the "evaluation of ML explanations [...] a multidisciplinary research" [ZJ21b, 15]. The inherent diversity in how explanations are interpreted further complicates their practical implementation [GR18, 2]. While general evaluation criteria exist, they must be adapted to different explainability methods and contexts. To address this, researchers attempt to summarize desirable features that can be generalized across multiple application scenarios [ZJ21b, 4]. The most suitable approach for a specific use case should be determined through comparison and iterative evaluation [ZJ21b, 2], as previously introduced for XAI method selection, which is based on XAI method evaluation.

Generally, the *IEEE Guide for an Architectural Framework for Explainable Artificial Intelligence* [Art24] presents a technological framework that provides measurable solutions for assessing explainability and outlines key requirements for generating human-understandable explanations across various use cases, such as healthcare and finance. We outline how our proposed XAI lifecycle stage design aligns with the framework in section 6.3. More generalizable perspectives on XAI method evaluation exist in the literature [RSB18] [SF19] [BP20] [DA21], complemented by research on XAI method-specific evaluation metrics, for example, as in [MN21] for feature-based methods. Ultimately, "[...] although quantitative proxy metrics are necessary for an objective assessment of explanation quality and a formal comparison of explanation methods, they should be complemented with human evaluation methods before employing AI systems in real-life" [MA21b, 9]. Human-centered metrics can be both quantitative and qualitative in nature [MA21b, 6], while functionality-grounded metrics are purely objective and quantitative [ZJ21b, 9].

The following introduces a combination of relevant evaluation criteria, aiming to highlight the complexity of evaluating explanations. In [SF19], the authors propose five global directions of thought for explanation evaluation:

- *Functional requirements* focus on understanding the structural setting surrounding the DNN model to select an appropriate XAI approach for the given problem [SF19, 2]. This step is crucial to identify interdependencies, since model quality impacts explanation quality.

- *Operational requirements* consider the end user’s interaction with the explanation to derive their required level of expertise for a seamless application [SF19, 3]. For this step, it is crucial to understand the intended use and domain of the intelligent system.

- *Usability requirements* define essential properties of explanations to ensure they are comprehensible and effective for users [SF19, 4].

- *Safety requirements* highlight the risk that XAI methods "[...] tend to reveal partial information about the data set used to train predictive models, these models' internal mechanics or parameters and their prediction boundaries" [SF19, 6]. The risk of exposing sensitive information emphasizes the need to manage disclosure carefully with respect to "[...] robustness, security and privacy aspects of predictive systems [...]" [SF19, 6]. For instance, in [SS24], the authors analyze the interaction between explanations and privacy-preserving DNN methods.

- *Validation requirements* stress the importance of thoroughly testing and evaluating explanations before deployment in real world applications [SF19, 7].

In the following, we introduce considerations to measure explanation validation (or quality) and usability requirements. The other considerations can be summarized as domain knowledge (operational), risk management (safety), and lifecycle-specific interdependencies (functional). They are more indirectly related to explanation evaluation but comprise necessary information for accurate design and evaluation interpretation.

The following criteria offer a foundation for quantitative XAI method evaluation metrics, while equally providing valuable guidelines for explanation design. First, we consider objective evaluation criteria to assess a "good" explanation based on [RSB18], who introduce nine criteria in a detailed and concise manner, including *Fidelity* and *Robustness*, for which we introduce a concrete evaluation method for LIME and SHAP explanations in form of a score [LGC24] in chapter 6:

- *Accuracy* refers to how well an explanation can generalize to new instances based on previously generated explanations [RSB18, 4]. This criteria can be complex to implement and monitor for specific use cases, since defining a fitting accuracy measurement strategy requires domain knowledge and a reasonable measure approach. As a result, even though accuracy is an important measure, it should not be the only evaluation assessment criteria. This tendency to define a multi-variate evaluation setup aligns with the previously introduced model performance evaluation procedure.

- *Fidelity* measures how well an explanation aligns with a model's prediction, essentially assessing how accurately the explanation reflects the decision-making process [RSB18, 4]. It reflects the "[...] extent to which extracted representations accurately capture the opaque models from which they were extracted" [BP20, 4]. As noted, "[i]n the case of an intrinsically interpretable model, fidelity is guaranteed by design" [MA21b, 5], which applies to models like decision trees but not to black-box systems. To ensure fidelity, an explanation must be both complete and sound, meaning it "[...] provides sufficient information to compute the output for a given input" [MA21b, 6] and remains reasonable within the model's domain-specific context [MA21b, 6]. High fidelity is crucial, as it ensures that explanations reflect the model's true decision-making process rather than spurious correlations. This, in turn, enhances human understanding and trust in the system [YF19, 3].

- *Consistency* evaluates how similar explanations are when generated by different models using the same data [RSB18, 4]. Ideally, explanations should be independent of the model choice and relate to the data. Given that both training and new data instances share the same application domain, explanations should be comprehensible within this context, while the underlying model remains interchangeable. The features identified as relevant should be consistent across models, reflecting their ability to capture meaningful patterns. If an explainability algorithm consistently highlights similar features for different models within the same application context, it can be considered reliable and more data-driven than model-dependent. In [Bre01], the author describes the phenomenon where different models, even highly divergent ones, achieve the same error rate on a test dataset as the *Rashomon Effect*[13]. He states, that "[...] there is often a multitude of different descriptions [equations f(x)] in a class of functions giving about the same minimum error rate" [Bre01, 206]. Applied to explainability evaluation, this suggests that if different models within a specific application context produce similar outputs, they should be explained in the same way by identifying similar relevant features. This would indicate that the explainer is consistent and relies more on the input data than on the model itself.

- *Stability*, or *Robustness* assesses how consistently explanations correspond to each other for similar input instances within the same model [RSB18, 4]. An explanation is considered more stable if small changes in the input data do not lead to significant variations in its output. Robustness requires the explanation to follow a clear internal logic, identifying relevant features in a way that remains understandable to humans. Ideally, explanations for similar instances should overlap proportionally, as the reasons for assigning a specific label should depend on the data's relevant features rather than the model itself. Some XAI methods may be more prone to respect this requirement than others. For instance, gradient-based explanation techniques are unlikely to exhibit high continuity due to noise [MG17, 10]. A high degree of robustness is desirable, as it suggests that a model generalizes well and will likely perform effectively on new data instances. Robustness is also

[13] Originally, the *Rashomon Effect* comes from a Japanese movie where four people experience the same event but describe it differently afterward [Bre01, 206]. Analogously, the event represents the sample, the people represent the models, and their differing stories correspond to the model outputs, which can be analyzed for deviations.

linked to clarity, a key interpretability feature, since an explanation should be "unambiguous" [MA21b, 6].

- *Comprehensibility* Comprehensibility (and interpretability) [GR18, 7], as introduced in section 2.2.2.1, refer to how well the target audience understands an explanation, [RSB18, 4] highlighting the model and explanation setup, and the respective human's level of knowledge. Closely related is transparency, which implies how well stakeholders grasp the ML model itself [Lip16, 4]. Since an explanation's primary goal is to ensure understanding, evaluating comprehensibility is essential for real world applicability. However, the level of technical detail required varies by audience: while end-users may need a simplified, non-technical explanation, data scientists can handle more complex details. As a result, understanding the respective target audience play a crucial role to design explanations [BP20, 4]. For an explanation to be comprehensible, it should avoid excessive complexity while still accurately reflecting the model's functionality, a balance also known as parsimony [MA21b, 6]. One way to assess comprehensibility is through human-machine task performance [ZJ21b, 8], but this is often subjective and difficult to quantify [ZJ21b, 9]. As previously mentioned, qualitative evaluation requires human test subjects to assess readability and usefulness, which can be resource-intensive [ZJ21b, 8]. Therefore, comprehensibility categorizes as human-centered evaluation [VL20, 43]. Ultimately, without comprehensibility, users cannot gain insights into a model's functionality, which is crucial for trust and AI integration into society.

- *Certainty* measures the alignment between a model's confidence in its prediction and how well this confidence is reflected in the explanation [RSB18, 4]. Ideally, an explanation should not contradict the model's confidence. If the model is certain about its prediction, the explanation should support this certainty. However, explanations can sometimes reveal that a model has learned the wrong function, even when it achieves high accuracy [LS19, 2]. In such cases, discrepancies arise between the model's confidence and the explanation, signaling potential issues in feature learning. Analyzing certainty helps identify inconsistencies between prediction confidence and explanatory justification, reducing the risk of misinterpretation. If an explanation convincingly demonstrates that the model has learned the correct features, the model's confidence is justified, leading to a high degree of certainty.

- *Novelty* is considered "[...] a form of certainty [...]" [RSB18, 4], as it examines the relationship between a model's confidence in its predictions and how well this is reflected in the explanation. However, unlike certainty, novelty specifically focuses on new instances that were not part of the training dataset. These instances should still fall within the general application domain but may differ from the training data distribution. As an instance deviates further from known examples, the model's confidence is expected to decrease, since previously learned relevant features may no longer apply. To meet the novelty criterion, explanations should explicitly convey this drop in model confidence, ensuring that the user is aware of the model's uncertainty in unfamiliar scenarios. As a result, XAI evaluation and model performance evaluation can be aligned.

- *Degree of Importance* evaluates how well an explanation reflects the significance of different features in a model's decision-making process, ensuring that relevant features influence predictions while non-relevant features do not [RSB18, 4]. This concept, also known as explanation selectivity [MG17, 7], can be assessed using feature relevance algorithms. One method involves sorting features by importance, progressively removing them, and analyzing the model's performance decline [MG17, 10]. The process concludes when removing features no longer impacts predictions, marked by low model performance evaluation metrics. Another approach [MN21] perturbs supposedly irrelevant features to see if the prediction changes. In the case of LIME, for example, all pixels except those deemed explanatory can be replaced with noise. The c-Eval score is then defined as the "[...] minimum distortion perturbation [...]" required to alter the model's prediction [MN21, 1].

- *Representativeness* evaluates how well an explanation reflects the model by determining how many data instances it covers [SF19, 5]. This criteria specifies the previously introduced XAI method *scope* in more detail, highlighting the alignment between quality evaluation criteria and method design. Explanations can be local (specific to a single instance), cohort-based (applicable to a subgroup), or global (describing the entire model) [SF19, 2]. If multiple new instances receive reasonable explanations following the same logic, cohort or even global representativeness may be inferred.

Two additional perspectives for evaluating XAI methods are *scalability* and *generality*. Scalability assesses how well the XAI method handles large input datasets and numerous weighted connections, ensuring it can maintain efficiency and accuracy as data complexity grows [BP20, 4]. Generality measures whether the selected XAI method can explain the entire underlying opaque model, or if there are limitations or necessary additions to fully capture the model's behavior [BP20, 4]. These metrics contribute to ensuring that XAI methods remain robust and applicable across different contexts and model complexities. Complementing the introduced explainability evaluation criteria, the following list provides an additional overview, as presented in [AS23, 7]. In addition, we align these criteria with three interpretability characteristics, namely *clarity*, *parsimony* [MA21b], and *broadness* [ZJ21b]. These perspectives highlight *comprehensibility* in more detail and provide a possible foundation for quantifiable criteria that lead to humanly interpretable explanations:

- *Quality* is essential, meaning explanations must be accurate and backed by evidence.[14] To specify this criteria, the previously introduced considerations provide relevant considerations. In [MA21b], the authors highlight explanation *fidelity* in particular. Accordingly, explanations need to be *complete* and *sound*, meaning they cover and explain the entire underlying model and are correct and truthful within the context of their specific application domain [MA21b, 3]. In cases when the explainer itself is another model (model-based), "[...] explanations provide sufficient information to compute the output for a given input, and thus always satisfy the completeness property" [MA21b, 6]. Ante-hoc methods always comply with the soundness property, while for post-hoc model-based explanations, soundness must be explicitly measured [MA21b, 6].

- *Quantity* of information should be balanced, providing enough information without overwhelming the recipient. This may include *clarity*, which means that the explanation is unambiguous and transfers a single meaning "[...] that is similar for similar instances" [MA21b, 3].

- *Relation* ensures that only pertinent details are included. This criteria aligns with the *parsimony* of an explanation, that evaluates how well it is presented and its level of complexity [MA21b, 3]. In terms of the human target audience, comprehensibility is key, with less complex explanations generally being more desirable to ensure better understanding.

- *Manner* focuses on how the explanation is delivered, emphasizing conciseness, structure, and avoiding ambiguity. These requirements equally provide valuable criteria regarding the *parsimony* of an explanation [MA21b, 3].

- *Context-oriented* design is crucial, as explanations should be tailored to different audiences, such as developers, regulators, and end-users, while possibly deriving elements for implementation from relevant domain knowledge. This information aligns with the *broadness* of an explanation, that refers to the range of potential application contexts in which it can be effectively used [ZJ21b, 3].

In summary, challenges remain, particularly in how to implement generalized requirements for a *good* explanation. The broadness of possible application domains makes generalizing explanations a difficult and time-consuming task, as each explanation needs to be tailored to its specific field. Simultaneously, this motivates methods for reliable explanation assessment to validate quality criteria in addition to resulting design guidelines. Further, adopting uniform terminology to reduce ambiguous definitions is a reasonable goal, and the previous sections aim to consolidate existing terminology. The next step involves the concrete in-

[14] This includes addressing AI lifecycle evolutions and existing interdependencies of the XAI method with the underlying data and model.

tegration of AI systems and their explanations into real world applications, with the actual benefits needing to be verified [MA21b, 3]. While XAI methods can be applied for a multitude of risk mitigating scenarios beyond user comprehensibility of AI outputs [Leb23] [DA21] [DS21], related risks [HA23] [SS24] emerge.

After emphasizing the complexities of design decision-making surrounding DL and explaining AI output, the following section summarizes AI pitfalls along the AI lifecycle development stage. These design-related intricacies comprise a common source for methodological errors. Their mitigation is supported through the application of best practices, or standardized procedures, and a high degree of AI literacy is required to implement RAI.

2.3. AI Pitfalls & Best Practices

Despite their promising performance, DNNs raise concerns among practitioners [HS20, 3], and may cause harm [DS21, 1]. As a consequence of the previously introduced intricacies surrounding relevant design considerations for DNNs, lifecycle design of AI systems comes with inherent risks and challenges. They arise from the technology's inherent dynamics and a continuously evolving real world, requiring a high level of AI literacy to navigate lifecycle effectively. This is equally recorded in the EU AI Act in Article 4 [Fut24].

In summary, intricate and interdependent design decisions span the entire AI lifecycle and, focusing on the development stage, this section introduces AI Pitfalls, i.e. methodological and conceptual errors that result in faulty system behavior, if not addressed. We start with a high-level perspective on the entire AI lifecycle in section 2.3.1. Next, model-related data design decisions are introduced in section 2.3.2. Sections 2.3.3 and 2.3.4 focus on model evaluation and explanation, respectively. Addressing these risk sources is crucial for RAI implementations, if not considered, they negatively impact the intelligent system's quality. Section 2.3.5 introduces relevant concepts surrounding reliable lifecycle design, including best practices, design pattern and standards. Overall, we emphasize the importance of adopting (generalizable) guidelines for AI lifecycle design, highlighting data, rigorous validation and explainability.

2.3.1. AI Lifecycle Design

The AI lifecycle comprises all necessary process steps from idea and system design until its decommissioning in the real world. The concrete implementation of underlying design decisions depends on individual use cases, as previously introduced for model development. We discuss the AI lifecycle in more detail in sections 3.1.3.1 and 4.2, where we propose a generic design. A combination of high-level process steps is generalizable across use cases. Broadly, lifecycle stages are composed of design decisions covering the data, model development, deployment, maintenance, and decommissioning. The AI lifecycle plays a crucial role in compliance assessment, which examines AI lifecycle process implementations [RA23]. However, not all stages and process steps are equally well researched, highlighting "[...] data collection, feasibility study, documentation, model monitoring, and model risk assessment" [HM21, 1]. Also, some lifecycle requirements, such as RM and documentation, cover the entire lifecycle with varying degrees of depth and expertise for different stages. As discussed in previous sections, design decisions related to DNNs are highly complex and require in-depth technical expertise. This complexity gives rise to potential AI pitfalls that must be effectively mitigated and evaluated.

Among the common reasons for AI system failures, the design of the AI lifecycle plays a central role. Failures often stem from errors of omission and commission in system design, as well as from an inadequate interpretation of input data. Additionally, even when the AI system itself is well-designed, performance issues may arise if the underlying hardware lacks the robustness to function reliably across different environments. [BC20, 1] Further, in [SD24], the authors analyze the origins of AI system failures, focusing on omission and commission errors in inputs, processing logic, and outputs. They identify 28 factors to trace AI failures and guide corrective actions. The "[...] frequency and severity of AI errors [...] in randomized controlled trials [...]" [KA24, 1] is assessed in [KA24] and the ongoing review examines AI medical devices as clinical interventions and analyzes performance errors, including subgroup-level outcomes. Further, in [RM21b], the authors evaluate different models for diagnosing and predicting COVID-19 from chest X-rays and CT scans, analyzing studies published between January and October 2020 to assess their potential clinical utility. Alarmingly, they find "[...] that none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases" [RM21b, 1]. Bias is an important topic, and in [Gic23], the authors review AI bias pitfalls and mitigation strategies in healthcare. They highlight bias as a continuum influenced by both human and machine factors and frame these pitfalls within the AI lifecycle.

Motivated by this diversity of possible AI pitfalls and their multi-faceted nature, we argue for organizing the AI lifecycle in a way that transparently represents all underlying processes in one place and in an iterative manner to serve as a practical tool for stakeholders to globally enhance AI quality. A reasonable step to support this endeavor is the establishment of a "theory of AI errors" [Bar24, 1] in relation to lifecycle design, which would support RM, equally highlighting the cognitive foundations of AI systems and their inherent limitations in adjusting to social and human contexts [Bar24, 1].

2.3.2. Data-related Challenges

Data forms the backbone of any AI system, serving as its primary connection to the real world. Therefore, ensuring data quality, representativeness, and integrity is essential when developing intelligent systems. However, aligning AI design with complex real world scenarios is a challenging task. These scenarios are shaped by diverse influences, including societal, governmental, and human factors, alongside the technical constraints of real world environments. For instance, the integration with existing software, infrastructures, and consideration of hardware limitations comprise valuable information for design decision-making. In section 5.1, we further outline data challenges related to medical data in particular. Generally, for data to be of high quality it must not only meet technical standards but also provide a comprehensive and accurate representation of the intended environment in all its dimensions.

Quality There is a "[...] a clear lack of research on potential data quality issues (e.g., ambiguous, extraneous values). These kinds of issues are latent in nature and thus often not obvious" [FH22, 1]. The authors introduce *data smells* as "[...] a counterpart to code smells in software engineering" [FH22, 1], and propose to organize data quality issues into context-*dependent* and *independent*. The latter is further divided into *obvious* problems such as missing values and duplicates, or *latent* challenges, including ambiguous values and intermingled data types [FH22, 3].

Deep Learning Further challenges arise during data preparation for DL, that directly impacts the model. For instance, *data leakage* should be avoided and *batch effects* respected. A batch effect arises when data from different sources are combined, and the class distribution samples vary significantly between those sources. For example, if malignant tumors are imaged with one MRI machine and benign tumors with another, a model might learn to distinguish tumors based on machine-specific differences rather than actual tumor characteristics. [MF22, 5]. Data leakage occurs when the assumption that the training data is independent of the evaluation data is violated [MF22, 3]. These examples result in models that inaccurately reflect the real world. Therefore, the method how to split the data into training, validation and test data, as well as when to perform preprocessing needs to be accurate [MF22, 8].

Bias Another important topic in the context of data is the multi-faceted concept of bias [ISO21a] [SR22]. For instance, imbalances in training data may result in biased predictions, affecting fairness and reliability if undetected. Or, in case of supervised learning, the annotator's biases may be reflected in their data labeling. Also, data from one institution may be biased. [Gic23, 2] We outline complexities surrounding bias and fairness in [EM25b].

Domain Domain-specific, or context-dependent pitfalls emerge, when real world information is inadequately translated for the model, or when there is a mismatch between the information provided by the data and the current reality. In [KC19], the authors present a thorough overview of key challenges regarding the adaptation of AI systems in healthcare, categorized as either ML-, implementation-specific, or regarding difficulties with the adoption or sociocultural barriers. For instance, AI models can inherit biases from historical medical records, leading to disparities in outcomes that should be interpreted in alignment with the use case at hand.

Finally, mitigation strategies comprise the application and study of high-quality guidelines such as best practices or standards. For instance, in [RH22], the authors provide guidance for designing the data engineering lifecycle. With respect to the comprehensive AI lifecycle and its evolutions, as well as the dynamic real world, implementing rigorous data quality checks and bias detection mechanisms is necessary. Also, ensuring diverse and representative datasets paves the way for RAI.

2.3.3. Evaluation & Metric Selection

Evaluating AI systems is not trivial, as the choice of performance metrics heavily depends on the structural setting, data distribution and domain knowledge. In chapter 7, we consider the risk *unreliable performance evaluation metrics* in more detail and outline a reliable design decision-making process for ECG multi-label classification. Overall, (medical) AI must adhere to regulatory and ethical standards that provide the foundation for defining what to measure beyond performance. This includes criteria such as fairness, robustness, or trustworthiness, and metrics provide one possible evaluation approach. Possibly, evaluation happens in combination with related design decisions, such as the data splitting strategy in case of cross-validation to assess model generalizability [SQ21].

Suitability Model evaluation in general can be hindered by using incorrect metrics for the use case at hand, resulting in misleading conclusions. Without additional validation, models risk learning inaccurate patterns, such as overfitting to training data, causing failures in their ability to generalize [RM22], which impacts their real world behavior. As a result, simply achieving a high accuracy with any model, although desirable, is not sufficient for performance assessment. In addition to implementing a metrics compilation, for instance, XAI methods provide means to analyze metrics' behavior in alignment with the model's learned features. Even if a model's training accuracy on tasks such as image classification is high, it might have learned the wrong features, making it non-generalizable [LS19].

Clever Hans Phenomenon The phenomenon where a model internalizes incorrect features while achieving high accuracy is similar to the *Clever Hans* phenomenon in psychology [LS19, 2].[15] One example involves the PASCAL Visual Object Classes 2007 dataset[16], where images of horses also contain a source tag. By generating a heat map of the explanation,[17] it becomes evident that the model has overfitted to the dataset and actually learned the source tag instead of the relevant features of a horse to classify the image [LS19, 3]. This example highlights the need for a multi-level evaluation strategy that should include testing of chosen evaluation metrics.

In summary, the complexity surrounding the question how to evaluate AI output results in the need to establish a comprehensive monitoring strategy that comprises a combination of fitting metrics to measure relevant tendencies from different angles under consideration of existing interdependencies. Globally, this would benefit from an organized approach to identify evaluation related AI pitfalls. For instance, common methodological pitfalls on metrics selection focusing on gastroenterology are introduced in [HSA22], where the authors analyzed existing studies. Further, they provide an open-source tool to support accurate metrics calculation.[18] Or, in [TJ21], the authors provide a detailed introduction to performance evaluation in healthcare settings. They consider "[...] why the performance measures are not 'exact'" [TJ21, 9], and identify the lack of gold standards, variability in expert diagnoses, and limited generalizability of data from single hospitals as key reasons. Considerations to implement the identified measurement strategy need to be aligned with (groups of) individual use cases, including relevant domain knowledge.

2.3.4. Explainability & Trust in AI

The application of XAI methods in itself can be described as a best practice for model development in general to avoid undesired outcomes in model behavior, such as the *Clever Hans Phenomenon*. Another example comprises on-going research on e.g. *model scheming* in the context of GenAI [MA25]. This concept refers to the hypothesis that AI agents may covertly pursue misaligned goals while hiding their true capabilities and objectives [MA25, 1], potentially resulting in undesirable model behavior. Therefore, as previously introduced, XAI methods provide valuable tools to shed light on the model's inner workings, which

[15] The *Clever Hans Phenomenon* was discovered in the early 20th century when a horse named Hans appeared to solve simple arithmetic by pointing to the correct answer. It was later revealed that the horse did not understand the calculations but had learned to interpret tiny movements in the interviewer's face, which influenced its behavior [SG13, 1].

[16] `http://host.robots.ox.ac.uk/pascal/VOC/`

[17] Creating a heat map allows data scientists to understand which features, in this case, which pixels, the model relied on most to make its prediction.

[18] `https://github.com/simula/medimetrics`

is important for human-model interactions along the AI lifecycle and to foster trust in general. However, the correct application of XAI methods is not trivial [HA23] [HS20].

Adaptability In addition to XAI methods' inherent complexities, they are impacted by the surrounding dynamics of the intelligent algorithm's intended environment. The real world evolves over time, involving new data or data drift: "The more dynamic an algorithm's context is, the more challenging XAI will be. These dynamics result in today's explanation becoming obsolete tomorrow" [dH22, 4], and the potential need for different explanations over time can result in confusion and a lack of public trust. Further, model explanations do not guarantee that the system operates based on the stated causality, as the true relationship between inputs and outputs may differ, especially, in the context of local explanations. [dH22, 4]

Privacy Other XAI method-related risks emerge based on the fact that explanations can inadvertently disclose information about both the model and the data, resulting in privacy concerns. As a result, conducting a thorough analysis of possible *information leakage* and exploring ways to mitigate associated threats is critical. Another concern is *explanation misuse*, which refers to the risk that an outsider, by analyzing multiple explanations, could extract extensive knowledge about the underlying model. This issue is particularly relevant when the model is subject to confidentiality agreements or proprietary constraints. Addressing and examining these risks in greater detail is crucial to ensuring that explanations do not compromise the security or integrity of the model. [SF19, 6]

In summary, not all explanations are useful and some may be misleading or lack relevance to end users or other stakeholders. To ensure effective explainability, AI systems should continuously monitor explanation quality for relevance and reliability. It is important, that explanations reflect the model's true decision-making process. Also, the interface between the intelligent system and its end-users requires special attention, highlighting user-centered design. An unfitting presentation of explanations may result in poor human-AI interaction. In chapter 6, we address the question how to design a risk-mitigated explanation lifecycle stage that is intended to be generalizable and result in quality by design.

2.3.5. Key Concepts for Reliable AI Design

Quality by design is a well-established principle in traditional software engineering, providing structured approaches to system reliability and robustness. This section introduces related key terminology. First, we explore the essence of *design decisions*, next, we define *best practices* and *design pattern* in the context of RAI. Finally, we broach the role of *standards*, which are introduced in section 3.3.1.2.

Design Decision A design decision involves choosing among multiple alternatives to determine how a system will meet its objectives. These decisions range from high-level architectural choices to low-level implementation details, significantly impacting performance, maintainability, and scalability. Thus, "[...] professional practice is a process of problem solving. Problems of choice or decision are solved through the selection, from available means, of the one best suited to establish ends" [SD86, 1]. In case of RAI, the "ends" comprise the intelligent system, which is rooted in lifecycle design and requires a dynamic approach. "Means" in this case are all existing realizations for design decisions, which comprise the in, the previous sections introduced, methods for model development. And "best suited" depends on a variety of considerations, including domain knowledge, technical skills, or subjective reasons. Overall, design decision-making is a non-trivial task, but critical to ensuring the desired and required quality. Further, in [MMH22], the authors introduce the notion *design decision competence*. While they focus on criteria for (end-)user integration in participatory design, the general hypothesis that design participants require competencies to qualify as "design decision-maker" is equally reasonable in case of RAI lifecycle design. In addition to AI literacy, identified competencies could include the ability to quickly adapt to evolving technological advancements, and competencies in working with multidisciplinary and diverse teams, for instance.

Best Practice Best Practices are general guidelines or proven methodologies that optimize software development processes.[19] For example, they include coding standards, security measures, and performance optimization techniques. In the context of RAI, best practices have a variety of different targets. For instance, in [RA24], the authors combine 46 best practices into an assessment framework for AI benchmarks. Other design principles address data engineering [RH22], AI user experience (UX) design for GenAI to foster a safe usage [WJ24], or focus on a particular use case scenario, as in [PA22a] for medical AI. Also, many best practices are found in form of e.g. blog articles or other publications [MH21]. Finally, compliance-invested institutions, such as the German Notified Bodies Alliance (IG-NB) equally publish guidelines for medical AI system design [RA23]. In [MH23], the authors conduct an analysis of how comprehensive best practices cover AI lifecycle design. The "[...] highest number of identified practices were model training and data cleaning" [MH23, 4], while "[...] model deployment, model monitoring, and data labeling are the ones with the lowest number of identified practices" [MH23, 4]. Overall, the need for applicable and accessible design guidelines is omnipresent, and best practices provide broad, experience-based recommendations for software quality and efficiency.

Design Pattern A Design Pattern is a general, reusable solution to a commonly occurring problem in software design: "In engineering disciplines, design patterns capture best practices and solutions to commonly occurring problems. They

[19] In a broader sense, best practices not only exist in the development context. For instance, in [BJ23], the authors introduce *Best Practices for Using AI When Writing Scientific Manuscripts*.

codify the knowledge and experience of experts into advice that all practitioners can follow" [LV20, 26]. Each pattern outlines a common problem, explores various potential solutions, discusses their trade-offs, and provides recommendations for selecting the best approach [LV20, 28]. In contrast, best practices are less organized, providing broad guidelines and less specified solutions. Design pattern serve as a structured approach to solving design challenges, improving maintainability, scalability, and overall quality. They are realized as templates that can be adapted to various situations. Thus, they offer valuable guidance to developers, potentially imparting essential design knowledge while also ensuring the quality of the final outcome. A comprehensive perspective on design pattern for ML is provided in [LV20], where the authors consolidate practical insights into the reasoning behind the tips and techniques used by experienced ML practitioners in real world applications [LV20, 16]. While introduced in 2020, the book continues to provide a valuable starting point. Focusing on how to design system stability, in [Yok19], the authors propose a new architectural pattern, aiming to enable easier failure analysis. Further, in [WH22], the authors propose software-engineering design patterns for ML applications. They focus on "[...] any patterns that include design structure directly or address design concerns of ML software systems indirectly" [WH22, 31]. In 2019, based on their search query, they "[...] identified 19 scholarly documents [...]" [WH22, 31]. This number highlights the fields novelty. More recent contributions, from 2024 emphasize the fields traction. For instance, in [RN24], the authors review existing design challenges, best practices, and key software architecture decisions in ML systems. Through a systematic literature review of 41 studies and 12 expert interviews across nine countries, the authors identify 35 challenges, 42 best practices, and 27 design decisions, highlighting their interconnections. Or, in [Ind24] the author focuses on the question how to "[...] integrate ML functionality into complex systems as architectural components" [Ind24, 1].

Standard Lastly, a Standard is a formally established set of guidelines that ensure consistency, quality, and interoperability across systems. Standards provide structured frameworks for best practices, ensuring compliance and reliability in software development. They explicitly address the interface between implementation and regulation, and their role in the context of RAI is introduced in more detail in section 3.3.1.2.

Focusing on model development, this chapter provides an overview of the rapid evolution and prevailing intricacies surrounding AI lifecycle design, highlighting DL. In light of the vast body of existing, possibly incomplete/incompletely documented design knowledge, we argue that a globally applicable and dynamic lifecycle planning blueprint is needed to consolidate design at the interface of implementation and regulation to support the on-boarding of RAI in the real world. The following chapters introduce MQG4AI, aiming to facilitate AI lifecycle design and assessment through structured IM. The next section 3.1 first defines RAI in more detail, which is ethical, lawful, and accountable [DRN23].

Part II.

METHODOLOGY BASED ON QUALITY GATES TOWARDS CERTIFIABLE AI IN MEDICINE (MQG4AI)

3 Context, Method, and Objective

After introducing the complexities surrounding model development, this chapter outlines our proposed methodology and its real world context in more detail. The next chapter 4 introduces MQG4AI's basic structure, and proposed conceptual building blocks to construct the AI lifecycle, and chapter 5 outlines two medical use cases in more detail, that form the foundation for the evaluation in part III.

We first define Responsible AI (RAI), which is lawful, ethical, and accountable [DRN23] in section 3.1. Next, we highlight the central role of Risk Management (RM), as a core component of AI Quality Management (QM), according to Article 17 of the AI Act [Fut24] in section 3.2. Section 3.3 introduces the lifecycle blueprint structure, which is illustrated on GitHub[1], before outlining vision and limitations in section 3.4.

3.1. Towards Implementing Responsible AI

Diaz-Rodriguez et al. present a holistic vision of RAI, integrating key concepts "[...] from ethical principles and AI ethics, to legislation and technical requirements" [DRN23, 2]. They "connect the dots", envisioning AI, whose decisions are "[...] accountable, legally compliant, and ethical" [DRN23, 18]. We agree with the authors that "[...] in order to realize trustworthy AI that is compliant with the law, we advocate for the development of RAI systems, i.e., systems that not only ensure responsible implementation meeting the requirements for trustworthy AI but also adhere to AI regulation" [DRN23, 19], aiming "[...] to attain [the] expected impact on the socioeconomic environment in which [the intelligent system] is applied" [DRN23, 8].

This section introduces our vision – towards implementing RAI (RAI) for Deep Neural Networks (DNN) in high-risk contexts. RAI systems, promote "[...] auditability and accountability during [their] design, development and use, according to specifications and the applicable regulation of the domain of practice in

[1] https://github.com/miriamelia/MQG4AI/blob/main/README.md

which the AI system is to be used" [DRN23, 18]. Therefore, we emphasize the role of the required AI Quality Management System (QMS) in Article 17 of the AI Act [Fut24], which "[...] is the foundational building block to ensure ongoing quality and compliance [...]" [AM24a, 2]. AI QMS realize the implementation of RAI along the AI lifecycle. Our focus lies on the developer, or provider-perspective at the interface with regulation, since it is their responsibility to prove the required quality of the intelligent system during compliance assessment.

The next section 3.2 highlights the central role of RM as part of AI QM. The interface of ethics, law, and implementation motivates our approach to design MQG4AI – a blueprint for comprehensive and continuous lifecycle planning through Information Management (IM), as introduced in section 3.3.

3.1.1. Lawful: Relevant Regulation

AI is a disruptive technology that is capable to change the world as we know it and, as of now, we have not yet exploited its full capacity for the multitude of possible application scenarios. Consequently, evaluating AI's global impact is not a trivial question, and we will need more time to fully grasp the technology's ramifications. Aiming to direct AI's effects into a desirable direction, *Regulation* exists. It is broadly defined as "an official rule or the act of controlling something" in the Cambridge Dictionary[2], and compliance with "applicable laws and regulations" [DRN23, 8] is crucial for RAI. This section introduces the complexity of relevant regulation for AI in medicine, focusing on the European regulation on AI, i.e. the EU AI Act, and briefly discusses the possible trade-off between regulation and innovation. The next section outlines the AI Act's ethical foundation.

Our main focus lies on the AI Act, which came into force on August 1st 2024. Other relevant regulation when implementing intelligent systems comprises horizontal regulation, such as the General Data Protection Regulation (GDPR), as well as vertical, or sector-specific regulation that may apply. This includes the Medical Device Regulation (MDR) (or the In Vitro Diagnostic Medical Device Regulation (IVDR)) for medical devices that incorporate AI.

GDPR The GDPR focuses on privacy and security through the establishment of rules regulating the processing and transfer of personal data of individuals within the EU.[3] Intelligent systems can not exist without data, and their quality is highly impacted by the underlying data it is trained on, therefore, the GDPR is a relevant regulation to consider during AI projects. This includes the alignment of complex

[2] `https://dictionary.cambridge.org/de/worterbuch/englisch/regulation`

[3] `https://www.consilium.europa.eu/en/policies/data-protection/data-protection-regulation/`

concepts that apply in both regulations, e.g. for *automated decision-making*, as analyzed in more detail in [Pal24], which may pose risks to the required *Human Oversight*, as defined in Article 14 [Fut24] or regarding the *right to an explanation*, as detailed in [MA24], which is outlined in Article 86 *Right to Explanation of Individual Decision-Making* of the AI Act [Fut24].

MDR/IVDR Currently, medical products are regulated according to the MDR/IVDR. A detailed legal analysis of both regulations and the AI Act (version draft April 2021) is presented in [WC22], outlining interoperability, as well as potential conflicts that are centered around "duplicative requirements" [WC22, 41], which need to be analyzed towards a "[...] harmonization of AI related standards under the MDR and IVDR [...]" [WC22, 41], which, before the AI Act came into force "[...] have been regulating the emerging AI systems in healthcare for years already [...]" [WC22, 13]. Overall, it is the European Commission's intention to avoid a "[...] double regulatory burden" [WC22, 2] between the horizontal and vertical regulation. Among others, identified, potentially critical areas cover inconsistent definitions for key terminology, the classification into risk classes and high-risk requirements, QMS procedures, notified bodies and conformity assessment, post-market surveillance, as well as how to address cybersecurity [WC22].

The EU AI Act The novel European regulation on AI envisages to implement a trustworthy integration of AI in society through regulatory requirements regarding the impact of AI in the real world. It establishes four risk levels based on the intended use and impact of intelligent systems on "health, safety, and fundamental rights" [Fut24] for quality assessment and control. Therefore, some intelligent systems are of unacceptable risk, such as "manipulative or deceptive techniques" [Fut24], and consequently prohibited in the EU. They are listed in Article 5 and centered around protecting an individual's fundamental rights. Medical intelligent systems are high-risk systems, as defined in Chapter III, Articles 6, 7, 57, and Annex I of the AI Act [Fut24], which is our focus. For such systems, specific regulatory requirements must be assessed in collaboration with an independent notified body, involving human stakeholders throughout the AI lifecycle, i.e. all processes and design decisions from the system's conceptualization to application in the real world, until its decommissioning. Providers must implement compliant systems through a comprehensive AI QMS, as outlined in Article 17, which is required to include a Risk Management System (RMS), as specified in Article 9 [Fut24]. Further, in Recital 1, the main objective to ensure a human-centric and trustworthy application of AI, which is based on the protection of fundamental rights in general, is highlighted. Protection for the environment, as well as AI's capability to positively impact current global challenges are summarized in Recital 68. [Fut24]

Regulatory Complexity & Innovation While these regulations are designed to protect fundamental rights, they may disproportionately impact economic growth by increasing time-to-market and operational costs. Implementing all regulatory requirements is already challenging for medical device manufacturers without considering the AI Act. With respect to the MDR, "[...] the key challenges [...] include additional workload for technical documentation, higher resource expenditure and cost increase, lack of clarity regarding regulatory requirements, and delays caused by a lack of availability of notified bodies" [AM24a, 4]. Aligning the AI Act with the MDR/IVDR has "[...] the potential to create a lot of additional work for manufacturers and health institutions [...]" [WC22, 13]. As a result of this complexity, difficulties regarding the implementation of regulation might negatively impact innovation, if not addressed. Especially for start-ups, regulatory requirements can be overwhelming due to limited resources, expertise, and funding. Further, notified bodies need to be equipped to assess AI quality, which requires comprehensive AI literacy for a successful integration of RAI in the real world [AM24a, 4]. The authors of [AM24a] discuss further existing challenges regarding compliant AI in medical products. From an official perspective, the EU AI Office[4], as the "[...] centre of AI expertise across the EU" is tasked with implementing the AI Act, "[...] fostering the development and use of trustworthy AI, and international cooperation", and it remains to be seen how this will unfold in the upcoming years.

Following an overview of the legal landscape surrounding (medical) AI, the next section explores the ethical foundation of the AI Act, an essential pillar of RAI systems designed to foster the implementation of Trustworthy AI (TAI), and "[...] AI systems should also adhere to ethical principles and values" [DRN23, 8].

3.1.2. Ethical: AI Trustworthiness

In Recital 14a, the EU AI Act states: "While the risk-based approach is the basis for a proportionate and effective set of binding rules, it is important to recall the 2019 Ethics Guidelines for Trustworthy AI", and emphasize that they "[...] should be translated, when possible, in the design and use of AI models", and incorporated in "the development of voluntary best practices and standards" [Fut24]. All AI within the EU therefore should strive to be a promoter of European ethics.

The AI Act's Ethical Foundation "Ethics Guidelines for Trustworthy AI", was introduced by a High-Level Expert Group established by the European Commission (HLEG) [Eur19], and is introduced in more detail in sections 3.2.2.2 and 3.3.5. As depicted in Figure 3.1, the proposed realization of TAI translates ethical principles that mirror fundamental rights into criteria and methods for trustworthiness assessment that need to be tailored to specific use cases for implementation.

Seven Key Requirements comprise *Human agency and oversight*, *Technical ro-*

[4] `https://digital-strategy.ec.europa.eu/en/policies/ai-office`

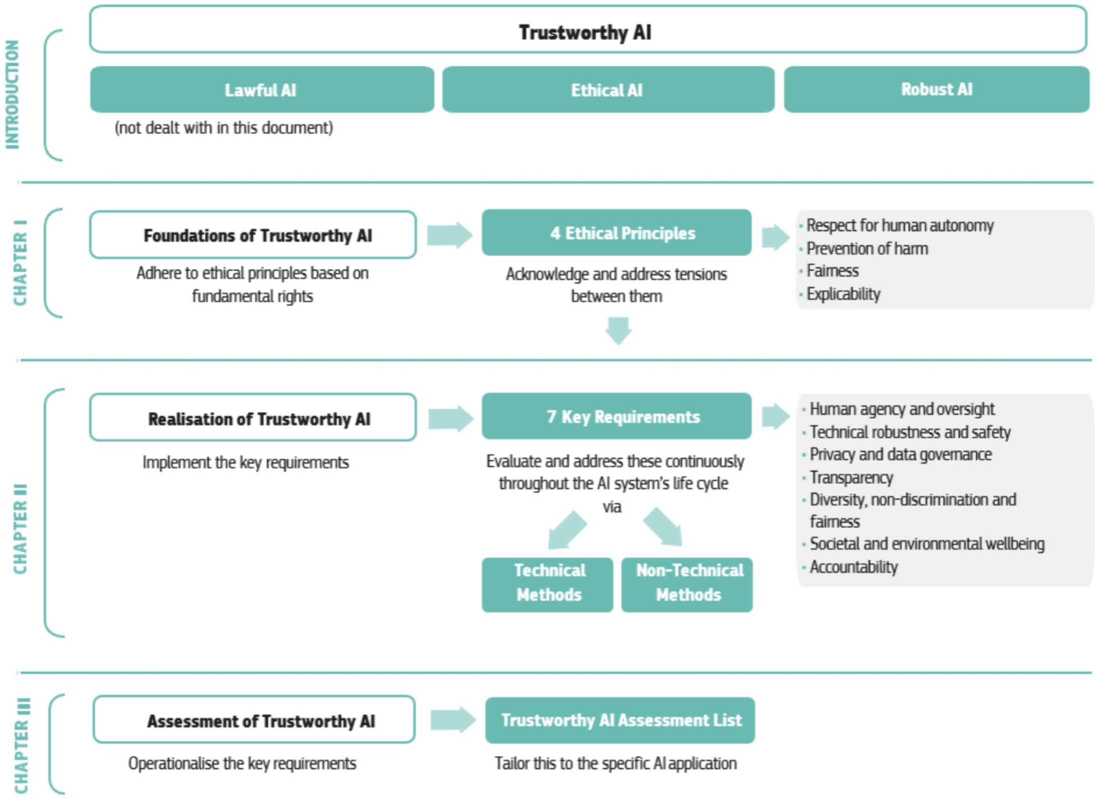

INTRODUCTION
Trustworthy AI
Lawful AI
Ethical AI
Robust AI
(not dealt with in this document)
CHAPTER I
Foundations of Trustworthy AI
Adhere to ethical principles based on fundamental rights
4 Ethical Principles
Acknowledge and address tensions between them
Respect for human autonomy
Prevention of harm
Fairness
Explicability
CHAPTER II
Realisation of Trustworthy AI
Implement the key requirements
7 Key Requirements
Evaluate and address these continuously throughout the AI system's life cycle via
Technical Methods
Non-Technical Methods
Human agency and oversight
Technical robustness and safety
Privacy and data governance
Transparency
Diversity, non-discrimination and fairness
Societal and environmental wellbeing
Accountability
CHAPTER III
Assessment of Trustworthy AI
Operationalise the key requirements
Trustworthy AI Assessment List
Tailor this to the specific AI application

Figure 3.1.: Ethical principles for trustworthy AI (TAI) and its realization, exactly as depicted in [Eur19, 8]. In contrast to TAI, RAI focuses on "[...] providing responsibility over AI products [...]" [DRN23, 18] for a practical implementation of TAI, which simultaneously adheres to regulation.

bustness and safety, *Privacy and data governance*, *Transparency*, *Diversity, non-discrimination and fairness*, *Societal and environmental well-being*, and *Accountability* that are intended to be addressed along the AI lifecycle through technical and non-technical methods. We outline the "Assessment List for Trustworthy AI" [Hig20], which provides a first step towards a more practical implementation of TAI in more detail in section 3.3.5, so that they can be tailored to individual use cases via AI risks, as introduced in section 3.2.2.2.

Trade-offs In general, it is to be emphasized that TAI criteria are interrelated and characterized by trade-offs, which need to be identified, and evaluated in alignment with individual use cases. For instance, regarding complex models, implementing explainability (transparency) may decrease accuracy (technical robustness and safety), and a use case-specific evaluation of that trade-off, which results in different risks, is required. In [van21] for instance, the authors consulted the general public for an opinion, which we find an interesting approach. They found that "[i]n healthcare scenarios, jurors favored accuracy over explainability, whereas in non-healthcare contexts they either valued explainability equally to, or more than, accuracy" [van21, 1]. However, even though the study provided jurors with expert sessions on complex topics surrounding AI, the question arises if the jurors were in fact capable to grasp related risks in their entirety. Overall, the evaluation of existing trade-offs is highly use case-specific and methods to address them need to be defined in a responsible manner on an individual level. On an abstract level, it should be a desirable outcome to establish standardized procedures for trade-off evaluation that are categorized for groups of use cases, such as the addition of the general public's opinion, for instance.

After presenting the AI Act's legal and ethical setting, the next section delves into approaches for the concrete implementation of RAI systems. This includes AI QM throughout the AI lifecycle and highlights the necessity for tools that bridge the gap between implementation and regulatory compliance.

3.1.3. Accountable: Compliant System Implementation

RAI systems need to ensure "[...] auditability and accountability during [their] design, development and use, according to specifications and the applicable regulation of the domain of practice in which the AI system is to be used" [DRN23, 18], so that ethical and legal AI enters the market. In addition to other methods, such as the establishment of regulatory sandboxes [Rus25] to test intelligent systems and their updates pre-market release from an official perspective [DRN23, 21], AI QMS [Fut24] support the implementation of legal and ethical AI. They offer a tool for provider to take responsibility for the intelligent system's behavior through closely monitoring all processes and design decisions that construct the individual AI lifecycle, addressing the interface with regulation.

This section first introduces the AI lifecycle, encompassing all processes and design decisions that constitute the intelligent system. It then highlights the AI standardization landscape, before providing a more detailed exploration of AI QM, emphasizing the need for flexible, customizable, and comprehensive tools that reflect the stages and complexities of the AI lifecycle.

3.1.3.1. The AI Lifecycle

ISO/IEC 5338:2023 Information technology — Artificial intelligence — AI system life cycle processes defines relevant procedures for AI lifecycle process management that are adapted from ISO/IEC/IEEE 15288 on system lifecycle processes and extended for AI-specific requirements, which are highlighted throughout the report [ISO23c]. In chapter 4, we analyze generic high-level lifecycle processes in more detail. Beyond the provider-perspective, the AI lifecycle bridges the gap to compliance assessment. For instance, the IG-NB, the German association of noti-

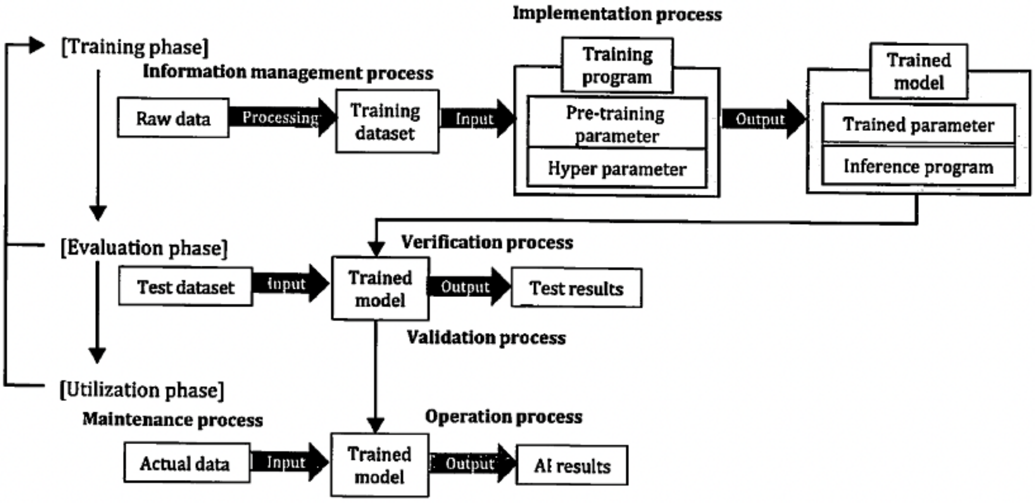


Figure 3.2.: The flow of AI-specific processes [ISO23c, 36].

fied bodies, base their guidelines on "[...] the idea that the safety of AI-based medical devices can only be achieved through a process-oriented approach, whereby all relevant processes and phases of the lifecycle must be considered" [RA23, 1].

AI Lifecycle Processes According to ISO 5338, basic AI lifecycle flow processes are defined as depicted in Figure 3.2: from data and use case comprehension over training, evaluating, and validating the model until maintenance and market-surveillance (or decommissioning). In response to AI's evolutionary character, the lifecycle processes are executed in an iterative manner. In addition, Figure 3.3 depicts the AI lifecycle flow integrated with related AI system lifecycle stages that start from inception, addressing contextual information before model design, and highlighting verification and validation procedures. These AI system lifecycle stages are aligned with *ISO/IEC 22989:2022 on Artificial intelligence concepts and terminology* [ISO22a, 36] and include relevant technical processes for each stage. The *training phase* corresponds to *design and development*, executing the implementation. During *evaluation*, *verification and validation* happens, and the *utilization* phase corresponds to *deployment*, where the transition into the real world happens, and *operating and monitoring* is conducted. The AI results are *continuously validated*. This cycle is interrupted either by e.g. updates such as new data, or quality divergence such as model drift that trigger a *re-evaluation*, or the system's *retirement*. In addition, this lifecycle functions as foundation for *ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system* [ISO23b]. Overall, the AI lifecycle comprises the most generalizable level across all types of AI. It is not trivial to implement and assess for all possible use cases, thanks to the technology's underlying dynamics, such as its non-deterministic and evolutionary character, as well as its opacity. The resulting system behavior in the real world is impacted by concrete design decisions that build individual lifecycles.

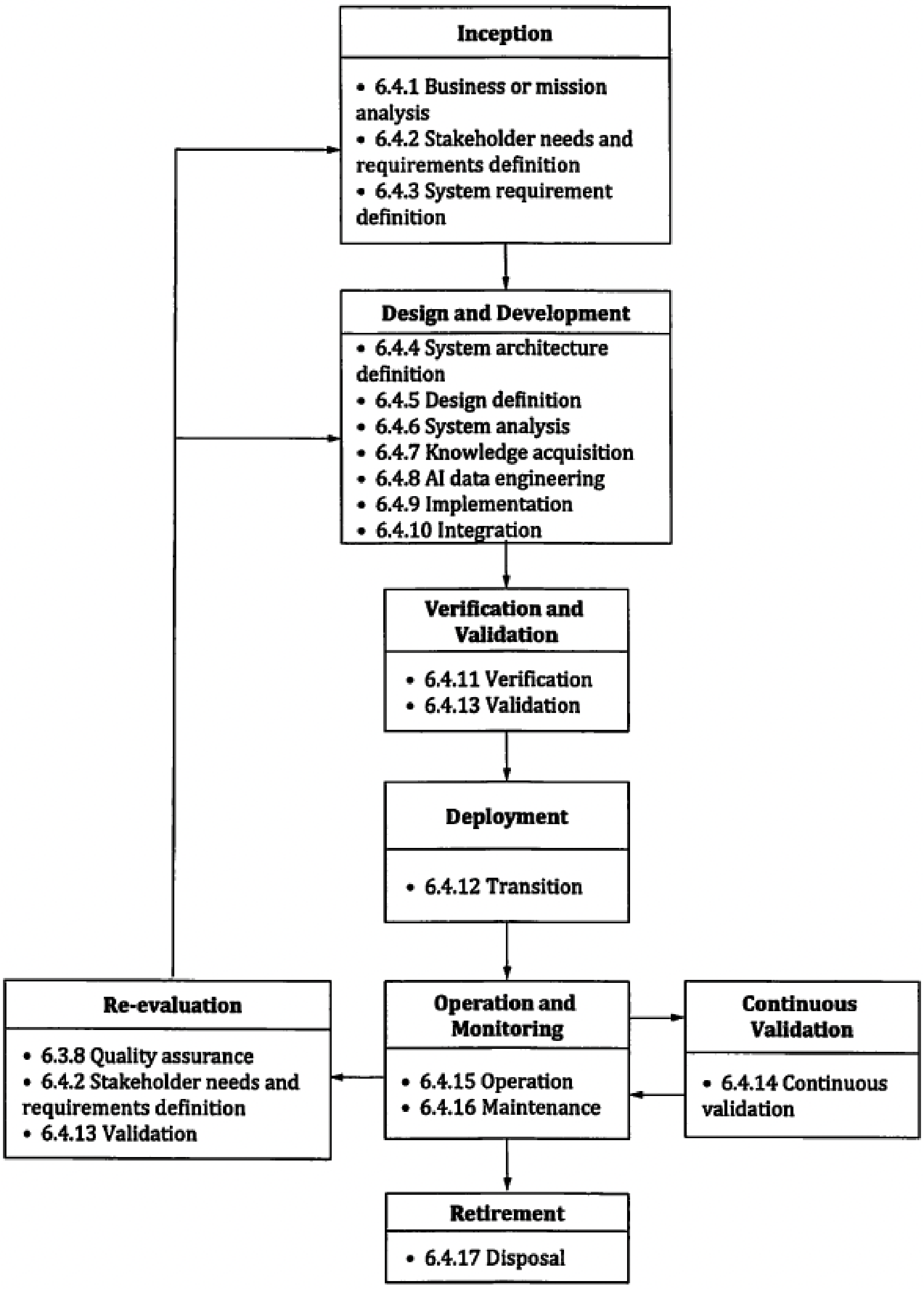

Inception
• 6.4.1 Business or mission analysis
• 6.4.2 Stakeholder needs and requirements definition
• 6.4.3 System requirement definition
Design and Development
• 6.4.4 System architecture definition
• 6.4.5 Design definition
• 6.4.6 System analysis
• 6.4.7 Knowledge acquisition
• 6.4.8 AI data engineering
• 6.4.9 Implementation
• 6.4.10 Integration
Verification and Validation
• 6.4.11 Verification
• 6.4.13 Validation
Deployment
• 6.4.12 Transition
Re-evaluation
• 6.3.8 Quality assurance
• 6.4.2 Stakeholder needs and requirements definition
• 6.4.13 Validation
Operation and Monitoring
• 6.4.15 Operation
• 6.4.16 Maintenance
Continuous Validation
• 6.4.14 Continuous validation
Retirement
• 6.4.17 Disposal

Figure 3.3.: AI system lifecycle stages including technical processes [ISO23c, 6].

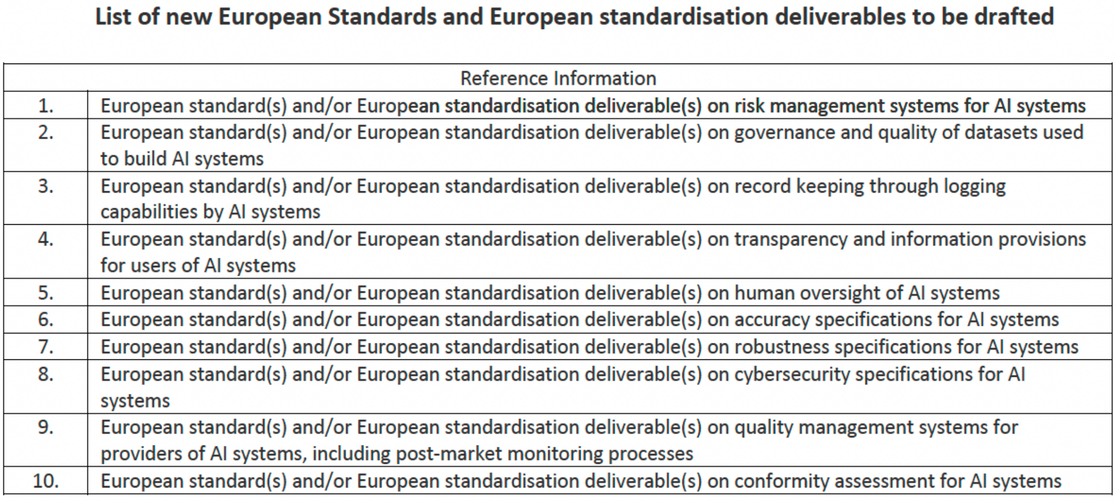
AI Literacy The intricacies of AI lifecycle design necessitate a comprehensive AI literacy of contributing stakeholders that construct and assess project-specific AI lifecycles. Fostering AI design knowledge among stakeholders is equally imposed by the AI Act in Article 4: "Providers and deployers of AI systems shall take measures to ensure, to their best extent, a sufficient level of AI literacy of their staff and other persons dealing with the operation and use of AI systems [...]" [Fut24].

3.1.3.2. The Role of Standards

Further supporting a high-quality implementation, standards play a crucial role to prove and evaluate a compliant implementation of lifecycle processes. They need to be aligned with the EU AI Act to be applicable in Europe, which is not always the case for existing international standards. European standardization for the AI Act can leverage international efforts with their standards being adopted and recognized within the European framework [SG24b, 2]. This section first introduces standardization bodies, highlights the state of European standards on AI next, before presenting a combination of examples on international and European level.

Standardization Bodies Standards are established at national, European, and international levels through designated organizations. Internationally, standards are developed by the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC). [Nat23, 11] With respect to the United States (US), for instance, the National Institute of Standards and Technology (NIST) develops standards and guidelines for organizations. At the European level, they are created by the European Committee for Standardization (CEN), European Committee for Electrotechnical Standardization (CENELEC), and the European Telecommunications Standards Institute (ETSI), collectively known as the European Standardization Organizations (ESOs) [Nat23, 11]. In addition, the Institute of Electrical & Electronics Engineers (IEEE) plays a prominent role in the field of artificial intelligence, with a particular focus on autonomous and intelligent systems [Nat23, 13]. Close cooperation between CEN and ISO, as well as CENELEC and IEC, fosters a high level of alignment between European and international standards. Priority is generally given to international standardization to leverage advantages for global trade and market harmonization. Under existing agreements between ISO and CEN, and IEC and CENELEC, certain standards developed at the international level can be adopted as European standards by CEN and CENELEC. [Nat23, 12]

European AI Standards The AI standardization landscape is in its early stages, and still under development, with most AI-specific publications being not much older than 2 – 3 years as of 2024. This leaves the concrete responsible implementation of individual compliant AI lifecycles an open question for the multitude of possible use cases. In general, "[t]he development of quality recommendations and standards [...] has to be a community-driven effort of many diverse stakeholders" [MH21, 120]. Figure 3.4 provides a summary on key directions for European AI standards, comprising RM, data quality and governance, record keeping, transparency, human oversight, accuracy, robustness, cybersecurity, QM, and conformity assessment. Note their similarity with the previously, in section 3.1.2 introduced criteria for TAI [Hig20], while the key areas put more emphasis on means to implement TAI criteria, which becomes visible for conformity assess-



List of new European Standards and European standardisation deliverables to be drafted

Reference Information	
1.	European standard(s) and/or European standardisation deliverable(s) on risk management systems for AI systems
2.	European standard(s) and/or European standardisation deliverable(s) on governance and quality of datasets used to build AI systems
3.	European standard(s) and/or European standardisation deliverable(s) on record keeping through logging capabilities by AI systems
4.	European standard(s) and/or European standardisation deliverable(s) on transparency and information provisions for users of AI systems
5.	European standard(s) and/or European standardisation deliverable(s) on human oversight of AI systems
6.	European standard(s) and/or European standardisation deliverable(s) on accuracy specifications for AI systems
7.	European standard(s) and/or European standardisation deliverable(s) on robustness specifications for AI systems
8.	European standard(s) and/or European standardisation deliverable(s) on cybersecurity specifications for AI systems
9.	European standard(s) and/or European standardisation deliverable(s) on quality management systems for providers of AI systems, including post-market monitoring processes
10.	European standard(s) and/or European standardisation deliverable(s) on conformity assessment for AI systems

Figure 3.4.: Relevant topics for European AI standards [Nat23, 26].

ment and QM. These areas comprise the TAI criteria fairness and societal well-being in an indirect manner through the assessment of criteria that need to be fulfilled by the intelligent system to achieve its intended purpose.

The European Commission published a document, where they introduce requirements for European standards, including how ISO standards that already exist within these 10 domains, are related to requirements of the AI Act, which will be applicable starting August 2026 [SG24b, 1]:

- *Risk Management*: Risk objectives and definitions need to be aligned to focus on product safety and not on risk to the organization, which is commonly followed by ISO/IEC. This is the case with RM for medical devices [ISO19], as will be outlined in section 3.2. Further, RM needs to focus on AI, and Article 9 in the AI Act defines related requirements, highlighting a continuous character and integrated testing procedures. Overall, standards do not need to prescribe specific risk treatment measures for every AI system, they must establish clear, explicit requirements concerning the processes and outcomes expected. Additionally, they should define key criteria and priorities that AI system providers must follow, particularly when evaluating and testing risk mitigation measures. [SG24b, 4]

- *Data Governance and Quality*: Article 10 in the AI Act outlines important aspects, which are not considered by ISO/IEC. Future European standardization needs to focus on identified risks, data quality metrics selection, and data governance along the AI lifecycle, among other criteria. [SG24b, 4]

- *Record Keeping*: Article 12 in the AI Act summarizes relevant requirements, and standards need to address logging, including "[...] clear requirements on how to establish a log-ging plan for AI systems" [SG24b, 5]. There are no

ISO/IEC standards mentioned. [SG24b, 4]

- *Transparency*: In Article 13, the AI Act lists required criteria, and related standards are expected to focus on information. Existing international standardization efforts could provide a robust foundation, particularly if the forthcoming ISO/IEC AI transparency taxonomy adequately addresses all transparency elements required by the AI Act, including those related to AI risks to individuals and society. [SG24b, 5] For instance, the in 2024 published *IEEE Guide for an Architectural Framework for Explainable Artificial Intelligence* [Art24] provides a comprehensive summary on relevant XAI concepts that can contribute to overall system transparency.

- *Human Oversight*: In Article 14, the AI Act defines related criteria, and standards should address means to support "[...] selecting, implementing and verifying the effectiveness of human oversight measures" [SG24b, 5]. ISO/IEC standards are not mentioned in particular, but a wide variety of oversight measures exist that ensure AI systems operate within intended constraints and allow natural persons to control or override their outputs when necessary. [SG24b, 5]

- *Accuracy*: Related standards should align with Article 15 of the AI Act. Standards should establish processes, methods, and techniques for reliably measuring accuracy and reporting it in accordance with best practices. It may not be feasible for AI standards to specify accuracy metrics and thresholds for every high-risk AI system or to detail measurement methods comprehensively. Nonetheless, several standards focusing on AI accuracy for specific system types, such as NLP or computer vision, are currently (as of 2024) being developed and will offer targeted requirements and guidance for certain applications. [SG24b, 5]

- *Robustness*: Article 15 of the AI Act provides requirements for robustness of the technical implementation. Some techniques for ensuring robustness in specific types of AI systems are beginning to be addressed by ISO/IEC standardization, providing a foundation for harmonized standards under the AI Act. However, it may not be feasible for standards to offer a comprehensive catalog of robustness techniques or prescribe measurement methods for every AI system. [SG24b, 6] Refer to item 5 in the list below for an example standard that is based on ISO.

- *Cybersecurity*: Article 15 in the AI Act outlines specific related requirements. Given the software-based nature of AI, some controls from existing standards, such as those in the ISO/IEC 27000 family, remain applicable, particularly for securing the IT infrastructure underlying AI systems. However, AI-specific vulnerabilities, such as data or model poisoning, model evasion, and confidentiality attacks, introduce new challenges. [SG24b, 6]

- *Quality Management*: AI QM is introduced in more detail in the next section and throughout this thesis, and Article 17 in the AI Act lists relevant requirements. Overall, existing QM standards for product safety legislation provide a useful reference, supporting compatibility with established processes in certain sectors, which we highlight for medicine in the next section. However, additional measures are necessary to address the unique characteristics of AI systems, whether integrated into physical products or delivered as software services. Existing international efforts, such as the ISO/IEC 42001 AI management system standard, offer relevant technical and organizational clauses. While not fully aligned with the objectives and approach of the AI Act, these can be referenced where appropriate in new QM standards, provided the focus remains on the specific risks and goals defined in the regulation. [SG24b, 6]

- *Conformity assessment*: Finally, standards should specify the procedures and processes required to assess the conformity of high-risk AI systems with the AI Act before they are placed on the market or put into service. Existing resources, such as the ISO CASCO toolbox[5], provide a foundation with generic principles and guidance for conformity assessment that can be leveraged in new standards tailored to the AI Act. New standards for AI conformity assessment should take a focused and practical approach, defining how conformity procedures, processes, and frameworks are applied and adapted for AI systems, particularly high-risk ones, in accordance with the legal requirements of the AI Act. Alignment between conformity assessment standards and the diverse requirements for high-risk AI systems is crucial. For instance, the assessment of the QMS is a key component of conformity under the AI Act. [SG24b, 6]

Examples In [OJ24], the authors publish a comprehensive overview of AI quality-related standards with a "[...] particular focus on the software aspects" [OJ24, 1], equally referencing the previously introduced AI lifecycle, concepts, and management standards [ISO23c] [ISO22a] [ISO23b]. In addition, focusing on the EU, the National Standards Authority of Ireland (NSAI), i.e. "[...] the Irish member body of CEN and CENELEC in Europe and ISO and IEC internationally" [Nat23, 11] for instance, published a comprehensive report on AI standards including their status [Nat23, 23]. The following list provides an excerpt of internationally published standards as of July 2023 (excluding Big Data-related publications) based on ISO [Nat23, 23], five of which include corresponding European standards in *italics*, as of December 2024:

1. ISO/IEC 25059:2023 Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems *EN ISO/IEC 25059:2024 Software engineering - Systems and software Quality*

[5] `https://casco.iso.org/toolbox.html`

Requirements and Evaluation (SQuaRE) - Quality model for AI systems

2. ISO/IEC 38507:2022 Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations

3. ISO/IEC TR 24372:2021 Information technology — Artificial intelligence (AI) — Overview of computational approaches for AI systems

4. ISO/IEC TR 24368:2022 Information technology — Artificial intelligence — Overview of ethical and societal concerns

5. ISO/IEC TR 24029-1:2021 Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview *UNE-CEN/CLC ISO/IEC/TR 24029-1:2024 Artificial Intelligence (AI) - Assessment of the robustness of neural networks - Part 1: Overview*

6. S.R. ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence

7. ISO/IEC TR 24027:2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making *UNE-CEN/CLC ISO/IEC/TR 24027:2024 Information technology - Artificial intelligence (AI) - Bias in AI systems and AI aided decision making*

8. ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) *I.S. EN ISO/IEC 23053:2023 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

9. ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology *ISO/IEC 22989:2023 Information technology - Artificial intelligence - Artificial intelligence concepts and terminology*

10. ISO/IEC TS 4213:2022 Information technology — Artificial intelligence — Assessment of machine learning classification performance

Another important standard is CEN/CLC/TR 17894:2024 "Artificial Intelligence - Artificial Intelligence Conformity Assessment". According to its official description, this document reviews current methods and practices for conformity assessment related to AI systems, covering products, services, processes, management systems, and more. It includes both horizontal and vertical industry perspectives, focusing on assessment processes, gap analysis, and defining objects of conformity for AI. Additionally, it examines challenges posed by AI in areas like software engineering and data quality, aligning with policy frameworks such as the EU AI strategy and standards from CEN and CENELEC member countries.

Finally, the document is aimed at technologists, standards bodies, regulators, and other stakeholders.

After introducing the current state of AI standards from a European viewpoint, the next section sheds more light on standardizable approaches to AI QM, considering sector-specific approaches for medicine.

3.1.3.3. AI Quality Management

In Article 17, the AI Act defines 13 requirements for AI QMS, which "[...] is the foundational building block to ensure ongoing quality and compliance [...]" [AM24a, 2]. They include a strategy for regulatory compliance; design, design control and verification; development, quality control and assurance; continuous examination, test, and validation procedures; technical specifications; data management; a post-market monitoring system (Article 72); reporting of serious incidents (Article 73); a risk management system (Article 9); communication with national competent authorities; record keeping; resource management; and an accountability framework. [Fut24] This section discusses the question how to implement AI QM, referencing AI standards, and emphasizes the need for concrete tools that foster a RAI integration, as well as facilitate the compliance process.

Implementation of AI QMS A general orientation on designing and integrating AI QM within an organization is outlined in ISO/IEC FDIS 42001 Information technology — Artificial intelligence — Management system [ISO23b], published in December 2023. With respect to European application, it currently comprises a preliminary standard for CEN. Preliminary[6] standards are still under development but planned. While the ISO version is described as being "[...] not aligned in objectives and approach with the AI Act [...]" [SG24b, 6], it identifies "[...] some relevant clauses at the technical and organizational levels. These could be referenced, as appropriate, by new standardization in quality management for AI, while ensuring that its focus remains on the specific risks and objectives captured in the legal text" [SG24b, 6]. In addition to addressing policy, resource, and documentation management, the standard equally emphasizes the importance of RM and the AI lifecycle [ISO23b]. Although the standard offers comprehensive conceptual guidance for addressing RAI, the concrete implementation of recommendations, such as considering "life cycle stages [...]; testing requirements and planned means for testing; human oversight requirements [...]" in processes for the responsible design and development of AI systems [ISO23b, 31], or "testing methodologies and tools; selection of test data and their representation of the intended domain of use; release criteria requirements" for AI system verification and validation [ISO23b, 32], remains relatively high-level. Overall, ISO 42001 builds on ISO/IEC 22989:2022 Information technology — Artificial intelligence

[6] `https://www.iec.ch/standards-development/stages`

— Artificial intelligence concepts and terminology [ISO23b, 1]. It equally references the previously, in section 3.1.3.1 introduced ISO/IEC 5338:2023 Information technology — Artificial intelligence — AI system life cycle processes [ISO23b, 32]. Further, EN ISO/IEC 25059:2024 Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Quality model for AI systems [ISO23a] provides a complementary overview to AI QM on relevant criteria that ensure the desired quality of AI systems through a software engineering lens, as depicted in Figure 3.5. In addition to similarities with key areas for European standardization, such as accuracy (performance-efficiency), robustness, transparency (usability) as well as security requirements, which are explained in section 3.1.3.2, they contribute requirements for a stable software implementation, such as compatibility, maintainability, and portability. The Fraunhofer institute continues EN ISO/IEC 25059:2024 and they propose an "AI System Product Quality Model for Safety-Critical Applications" [HL24, 11]. They break "[...] down high-level requirements into verifiable properties of AI systems" [HL24, 12] and include the use of verification trees to illustrate the fulfillment of quality criteria. This approach provides a useful means to translate AI Act-compliant requirements into AI lifecycle design decisions, similar to the concept of 'compliant best practices'.

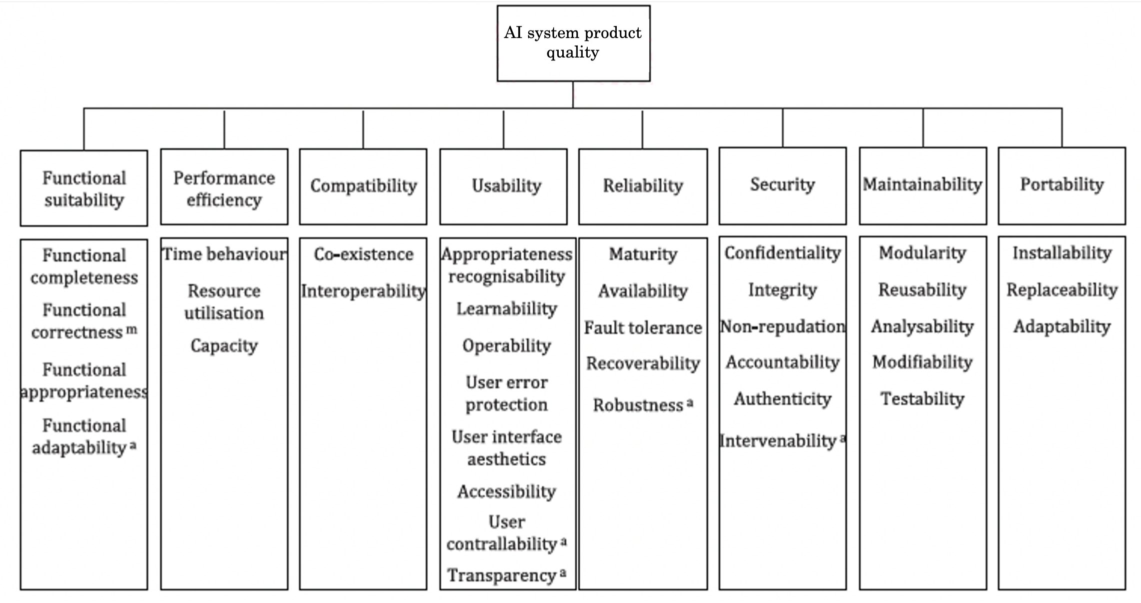


Figure 3.5.: Quality Criteria of AI system in software engineering terminology, based on EN ISO/IEC 25059:2024 [ISO23a, 3].

In addition to standards, various published frameworks and approaches contribute partial knowledge to the implementation of AI QM requirements, ranging from documentation management and (domain-specific) evaluation methods to software requirements in the AI context and the application of ethical principles [Gol24] [FL22] [RS21] [San20] [HA22]. For instance, explicitly addressing AI QM, in [ML24], the authors focus on the RM component of AI QM, addressing model integration and application in the clinical context. They emphasize that in addition to developers, who should follow best practices and clearly communicate how system robustness is ensured, "[...] the users are responsible for implementing the AI system in a QM system and setting up appropriate [quality assessment] measures" [ML24, 350] within the intended clinical workflow. Drawing on experiences from radiation oncology [ML24, 345], they outline how to set up AI QM that mitigates identified risks [ML24, 346] based on "[...] process mapping, failure mode and effect analysis (FMEA) and fault trees (FT)" [ML24, 345].

Tools for AI QMS Meanwhile, the use case-specific implementation of AI QMS criteria [Fut24], which must address the technology's evolutionary and non-deterministic nature, remains a significant challenge. Due to the complexity of the topic, and compared to the volume of research on AI risks and the reliable evaluation of individual AI components, "[...] there is less research on designing and implementing a quality management system (QMS), [...] and no clear suggestion exists [...]" [MRM24, 2]. Consequently, there is a pressing need for comprehensive "practical tools" [MRM24, 1] that support AI QMS processes and requirements beyond the scope of scientific prototypes [San20, 11]. These tools are essential to facilitate the implementation of the AI Act, promote AI innovation, and contribute to the responsibilities of the AI Office [AM24a, 4]. As a result, AI governance start-ups are emerging, like Munich-based *trail*[7] for instance, which provides an AI Governance Co-pilot. As noted, "the integration of [AI QMS] across the life cycle to provide a holistic system is critical for wide scale use" [San20, 11], making lifecycle-adaptation the recommended approach. In [MRM24], the authors propose a modularized, microservices-based AI QMS tool aimed at enabling user-guided quality checks of AI system components throughout the AI lifecycle through a modularized setup. Their prototype focuses on automating RM for LLMs and generating the necessary documentation to support compliance. To the best of our knowledge, few publications currently exist on AI QMS tools, and further contributions can be expected in the future.

Non-AI QMS Procedures Finally, tools for AI QM should provide means that allow for integration with respect to existing, sector-specific QM procedures, to facilitate a seamless transition for AI quality assessment on both the regulatory and provider sides. With respect to the medical domain, "[...] it is not clear how the essential world-wide standard for medical devices QMS ISO 13485:2016 will relate to the draft AI Act. This standard has a large (but not complete) overlap with the QMS requirements under the MDR and IVDR" [WC22, 31]. In addition to management responsibility, such as the definition of quality objectives and policy, or the definition of responsibilities, ISO/IEC 13485 highlights resource and continuous documentation management for medical devices, including RM. Further, focusing on the medical context without particular emphasis on AI, it underlines qualitative planning, design and development, as well as deploying the product into production, which includes e.g. "[i]nstallation activities" [ISO16,

[7] `https://www.trail-ml.com/`

18]. Design and development of medical devices concretizes requirements such as planning of relevant stages, the definition of product criteria, which, among others include "[...] performance, usability and safety requirements according to the intended use" [ISO16, 14], quality evaluation of generated outputs, as well as verifying and validating the selected processes and design decisions to achieve the desired quality [ISO16, 15]. Finally, means to monitor the performance of the QMS such as feedback on meeting "customer requirements" [ISO16, 22] and how to approach non-conformity need to be put into place. [ISO16] On a global level, the described steps seem to be alignable with AI QM, and the AI lifecycle, as introduced earlier, and "[...] it is expected that medical device manufacturers will continue using the ISO/IEC 13485 QMS and incorporate the requirements of Art. 17 within their existing medical device QMS functions" [AM24a, 2], while the question remains how exactly this will be realized.

In summary, AI QMS should provide flexibility to correspond with the complex technological dynamics along the AI lifecycle, closing the gap to requirements of individual use cases. Simultaneously, they should be alignable with sector-specific regulation for a seamless integration with existing processes. This facilitates the already complex certification process for high-risk AI, and fosters a responsible innovation. Therefore, to accommodate the diverse range of domains that integrate AI, the implementation of generic and customizable AI QMS software offers valuable support, and needs attention.

After outlining the legal and ethical landscape surrounding AI, as well as introducing the role of standards, and AI QMS as a means to implement RAI, including challenges regarding the concrete implementation of AI QM, the next section underlines AI RM – a central component of AI QMS.

3.2. The Central Role of Risk Management

This section highlights the central role of RM towards implementing RAI. As introduced in section 3.1, RM is required by law, as part of AI QMS. We start with outlining conventional RM processes, emphasizing the use case-specificity of their realization during individual projects. They are transferred to AI on an abstract level, as implied in Article 9 of the AI Act [Fut24]. Aligning with the technique's opaque, stochastic and evolutionary nature, "[t]he RMS shall be understood as a continuous iterative process planned and run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic review and updating" [Fut24]. Next, we introduce AI-specific RM, focusing on AI trustworthiness, which comprises the basis for a proposed information block of the, in the next section 3.3 introduced MQG4AI blueprint towards RAI lifecycle planning.

3.2.1. The Risk Management Process

This section introduces RM sub-processes based on ISO 14971:2019 on the "Application of Risk Management to Medical Devices" [ISO19], as well as risk concepts, as identified in the AI Risk Ontology (AIRO) [GD22][8], which comprises the AI Act and ISO RM standards [GD22]. They are globally aligned, as we illustrate. ISO 14971:2019 is a, with the MDR harmonized standard, which in turn is harmonized with the AI Act, as postulated in Annex I [Fut24]. This qualifies ISO 14971 as orientation for AI Act-conform RM, since "[h]armonised standards for other EU product safety legis-lation can provide a reference, as standards for the AI Act will also be product-oriented" [SG24b, 4]. The standard defines conventional RM for medical devices, and ISO/TR 24971 summarizes practical guidelines how to implement the standard. The conventional medical RM process, as depicted in Figure 3.6, comprises *risk analysis*, *risk evaluation*, and *risk control* measures in alignment with the system's intended purpose and foreseeable misuse. This abstract approach is transferred to AI and related risks, for "[...] those which may be reasonably mitigated or eliminated through the development or design of the high-risk AI system, or the provision of adequate technical information", as stated in Article 9 of the AI Act [Fut24].

3.2.1.1. Risk Analysis

Risk Analysis comprises the identification and estimation of risks. The AI Act defines *Risk* as "[...] the combination of the probability of an occurrence of harm and the severity of that harm" in Article 3 [Fut24], which concurs with ISO 14791 [ISO19, 5]. Consequently, *risk sources* need to be identified, *risk events* derived and their occurrence estimated to achieve system *safety*, which is defined as "freedom from unacceptable risk" in ISO 14791 [ISO19, 6].

Prerequisites: Identifying and managing risks requires alignment with the specific application and stakeholders. Therefore, parameters describing the system's function in the real world, such as the system's intended purpose and area of impact, as well as relevant stakeholder roles need to be identified. They are introduced in more detail in sections 3.3.3.1, and 3.3.3.5. In addition, reasonably foreseeable misuse scenarios need to be outlined. Misuse entails scenarios deviating from the intended purpose. Identifying these scenarios is essential to anticipate unintended application events that may result in risks. With respect to AI, *reasonably foreseeable misuse* is defined as "[...] the use of an AI system in a way that is not in accordance with its intended purpose, but which may result from reasonably foreseeable human behaviour or interaction with other systems, including other AI systems" in the AI Act in Article 3 [Fut24].

[8] https://delaramglp.github.io/airo/

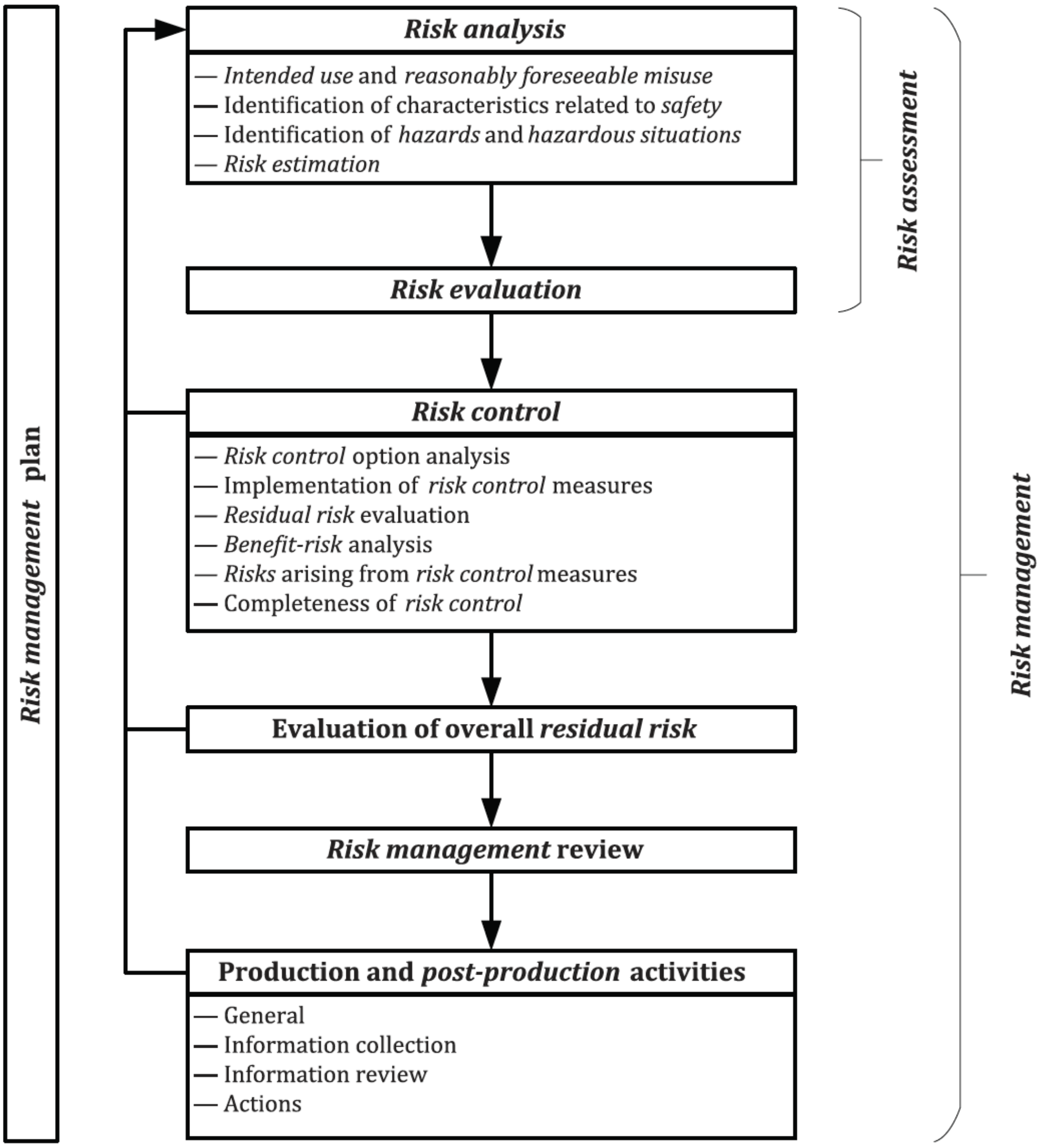

Risk management plan
Risk analysis
— Intended use and reasonably foreseeable misuse
— Identification of characteristics related to safety
— Identification of hazards and hazardous situations
— Risk estimation
Risk evaluation
Risk control
— Risk control option analysis
— Implementation of risk control measures
— Residual risk evaluation
— Benefit-risk analysis
— Risks arising from risk control measures
— Completeness of risk control
Evaluation of overall residual risk
Risk management review
Production and post-production activities
— General
— Information collection
— Information review
— Actions
Risk assessment
Risk management

Figure 3.6.: Risk Management Process [ISO19, 8].

Risk Sources: Risk sources are derived based on the identified system-specific prerequisites approximating and continuously monitoring the system's behavior in the real world. ISO 14971 defines *hazard* as a "potential source of harm" [ISO19, 2], and *harm* as an "injury or damage to the health of people, or damage to property or the environment" [ISO19, 2]. The AIRO[9] risk concepts provide a slightly more detailed perspective on potential sources of harm:

1. A *hazard* is analogously to ISO 14971 defined as a "source of potential harm".

[9] https://delaramglp.github.io/airo/

Hazards can be interpreted to arise from conditions or flaws, such as biased datasets.

2. A *threat* is defined as a "potential source of danger, harm, or other undesirable outcome", which we interpret as a more active perspective on the creation of harm, such as cyber attacks.

3. A *vulnerability* "[r]efers to properties of an entity, e.g. AI system or AI component, resulting in susceptibility to a risk source", which describe system weaknesses or gaps that increase exposure to hazards or threats, or, characteristics that could affect *safety*, as formulated more broadly in ISO 14971 [ISO19, 6].

Events are incidents triggered by risk sources, leading to potential harm, which when estimated regarding severity and likelihood result in risks. A single risk source may cause sequences of events. ISO 14971 introduces the term *hazardous situation*, that describes a "circumstance in which people, property, or the environment is/are exposed to one or more hazards" [ISO19, 2], which equally result from identified risk sources.

Likelihood is defined as the "[c]hance of an event happening" according to the AIRO risk concepts, and refers to the probability of an event occurring, considering risk sources and contextual factors. In ISO 14971 this is called "probability of occurrence of harm" [ISO19, 5].

Severity "[i]ndicates [the] level of severity of an event that reflects [the] level of potential harm" according to the AIRO risk concepts. In ISO 14971 wording, severity measures "possible consequences of a hazard" [ISO19, 6]. For this analysis, the AIRO risk concepts introduce two related terms:

1. *Consequence* is defined as the "[o]utcome of an event affecting objectives", for instance discrimination of minority groups or system failure, when the high-risk system is intended to adhere to TAI criteria such as non-discrimination and fairness, or robustness and safety. Consequences lead to impacts.

2. *Impact* is defined as the "[o]utcome of a consequence on persons, groups, facilities, environment, etc.", which identifies use case-specific realizations of identified consequences, such as effects of system failure, or concrete results of a possible discrimination of minority groups, e.g. women being less likely diagnosed correctly because of under representation in a data set, for instance. Impacts affect specific areas, in accordance with the system's previously identified *area of impact*. They also directly affect AI subjects, which describe individuals that interact with or are affected by the system.

In summary, during risk analysis, hazards that can lead to multiple hazardous situations through foreseeable sequences of events, which in turn can cause multiple harms need to be defined and estimated. Based on the combination of severity and likelihood of those possible harms, risks are identified, including AI-specific risks for intelligent medical products, as explored in the next section.

3.2.1.2. Risk Evaluation

Risk Evaluation is based on the previously executed risk analysis, and "[f]or each identified hazardous situation, the manufacturer [or provider] shall evaluate the estimated risks and determine if the risk is acceptable or not, using the criteria for risk acceptability defined in the [RM] plan" [ISO19, 12], according to ISO 14971.

Iterative Character: Highlighting the iterative character of RM, risk evaluation includes assessing *residual risk*, which remains after risk control measures have been implemented, as well as novel identified risks that may result from implemented risk control measures. This process repeats itself until the remaining (residual) risk is evaluated as acceptable. Then, risk control measures result in "information for safety" [ISO19, 13], or transparent communication to the user, as an example for the lowest possible mitigation measure. Consequently, the user should possess the level of knowledge that is necessary to understand the message. Otherwise, further risk controls need to be implemented, and the process restarts.

Benefit: In Annex 1, Chapter 1, the MDR promotes "[...] the reduction of risks as far as possible without adversely affecting the benefit-risk ratio" [Eur17]. ISO 14971 defines *benefit* as "positive impact or desirable outcome of the use of a medical device on the health of an individual, or a positive impact on patient management or public health" [ISO19, 2]. A *benefit-risk analysis* [ISO19, 14] can support the evaluation of residual risk, if "[...] it is not judged acceptable [...] and further risk control is not practicable" [ISO19, 14] through comparing "[...] the benefits of the intended use [...]" [ISO19, 14] with the residual risk.

With respect to AI, the AI Act highlights RMS maintenance in Article 9 [Fut24], which requires post-market release monitoring of identified risks, and implemented risks controls, which may result in a re-evaluation and updates during production. As a result, kick-starting the RM process is possible along the complete AI lifecycle, and needs to be considered from the beginning of conceptualization.

3.2.1.3. Risk Control

Risk Control is defined as the "process in which decisions are made and measures implemented by which risks are reduced, or maintained within specific levels" in ISO 14971 [ISO19, 5]. Similarly, the AIRO risk concept describe a risk control as "[a] measure that maintains and/or modifies risk (and risk concepts)". Consequently, risk control measures can address all components of the previously introduced risk analysis process, influencing forms of risk sources, and mitigating risk enabling events to either affect their likelihood, or severity, or both.

Forms: In Annex 1, Chapter 1, the MDR states that risks need to be eliminated or reduced "[...] as far as possible through safe design and manufacture" [Eur17]. According to ISO 14971, options for risk control include inherently safe design, protective measures, and information for safety, as well as training to users if applicable [ISO19, 12]. Focusing on AI and the technique's complexity, as well as our current state of AI design knowledge, the AI Act highlights the role of testing in Article 9: "High-risk AI systems shall be tested for the purpose of identifying the most appropriate and targeted [RM] measures. Testing shall ensure that high-risk AI systems perform consistently for their intended purpose and that they are in compliance [...]", and "[t]he testing of high-risk AI systems shall be performed, as appropriate, at any time throughout the development process, and, in any event, prior to their being placed on the market or put into service. Testing shall be carried out against prior defined metrics and probabilistic thresholds that are appropriate to the intended purpose of the high-risk AI system" [Fut24].

Verification: Continuous testing of AI RM aligns with the requirement, that risk control measures need to be verified. This is defined as "confirmation, through the provision of objective evidence, that specified requirements have been fulfilled" [ISO19, 6], while objective evidence describes "data supporting the existence or verity of something" [ISO19, 4]. This includes verifying the implementation, and effectiveness of risk controls, which is part of the applied QMS [ISO19, 13]. ISO 13485:2016 on "Medical devices — Quality management systems — Requirements for regulatory purposes" [ISO16] offers guidance on the implementation of design and development verification: "The organization shall document verification plans that include methods, acceptance criteria and, as appropriate, statistical techniques with rationale for sample size" [ISO16, 15].

Documentation: The comprehensive RM plan must be documented to ensure transparency and traceability. ISO 14971 describes a *RM file* as a "set of records and other documents that are produced by [RM]" [ISO19, 6], which is the foundation for compliance assessment.

After introducing fundamental concepts related to RM, the following exemplifies the risk *unreliable performance evaluation metrics* for AI systems. The hazard of incorrect measurement information as mentioned by ISO 14971 can be transferred directly to performance evaluation metrics for intelligent systems. The sequence of events includes wrong measurements that are not detected by the human stakeholder and result in the hazardous information that may result in misdiagnosis by the clinical user based on incorrect information, which can harm the patient. [ISO19, 35] As a result, reliable performance evaluation metrics, as well as their reliable interpretation are crucial for risk control. Generally, selecting and interpreting performance metrics depends on the respective use case and is prone to evolving over time, which implies a scientifically-based evaluation of suitable metrics with respect to the individual data set composition and tuning objective, as well as a monitoring strategy adapted to different stakeholder views. Finally, the use case-adapted performance evaluation concept needs to be documented, including the identification of interdependencies that might trigger changes after to-market release. Chapter 7 introduces a possible risk control in form of a metrics compilation selection process, which comprises a combination of design decisions along the AI lifecycle. This approach is interpreted as a design verification procedure included with the MQG4AI blueprint through overall lifecycle planning, which is introduced in section 3.3.

3.2.1.4. Contextual & Iterative Execution

As of now, there is no medical device industry standard to estimate risks, and "[f]unctional safety and its design concepts, as applied in other industries, have so far found little application in the field of medical technology" [PB21, 1]. Consequently, risk assessment (analysis and evaluation) is carried out on an individual use case-specific basis based on the previously introduced concepts. In ISO 14971, *risk estimation* describes the "process used to assign values to the probability of occurrence of harm and the severity of that harm" [ISO19, 5], as foundation for *risk evaluation*. As a general guideline, a risk matrix contrasting severity of harm with its probability of occurrence is recommended by notified bodies such as the British Standards Institution (BSI) Group [van20, 7] or the Johner Institute [Joh]. However, this approach does not cover all requirements, thus only contributes towards comprehensive RMS. Levels of severity and occurrence need to be adapted to the respective risk [van20, 8], which leaves decisions to the humans in charge, who adapt their interpretation to the respective use case. Also, a comprehensive approach to risk identification is necessary so that risks are not overlooked. This openness towards individual, and situation-specific human decision-making poses challenges for compliance assessment in the long-term.

Further, the AI-part adds additional risks that need to be addressed, stemming from the technique's inherent dynamics, such as its opacity and evolutionary character, as well as use case-specificity of reliable design choices. Consequently, as pointed out by ISO 5338 on AI system lifecycle processes [ISO23c], and in Article 9 of the AI Act [Fut24], this results in an ongoing and iterative RM process that needs to be considered during the complete AI lifecycle, and documented in alignment with the evolving AI system. The next section explores RM for AI, and how to classify risks, aiming to support a comprehensive and organized approach to risk analysis.

3.2.2. AI-Specific Risk Management

Intelligent systems contribute novel requirements for their reliable real world integration. These are mainly centered around the capability of DNNs to continuously learn, their broad applicability, stochastic nature, as well as inherent opacity for human understandability. Addressing this complexity within the context of individual projects poses risks, which need to be managed.

AI standards once published, and aligned with European regulation, as introduced in the previous section 3.1.3.2, provide a means to execute RM processes. Generally, "[...] standardisation should explicitly take into account the risks identified as part of the [RM] process" [SG24b, 4], as underlined by the European Commission. RM comprises a key area for the EU standardization landscape [SG24b, 4]. EN ISO/IEC 23894:2024 *Artificial intelligence – Guidance on risk management* is a recently published international standard that is aligned with the RM processes, as previously introduced and concentrates on how to integrate AI RM processes within the organization, as well as highlights mapping risks along the AI lifecycle. It extends ISO 31000 on *Risk Management - Guidelines* with AI-specifics focusing on impacts to human stakeholders, the inclusion of affected stakeholders, and human oversight. "Likewise, it recommends identifying how AI systems or components interact with preexisting societal patterns that can lead to impacts on equitable outcomes, privacy, freedom of expression, fairness, safety, security, employment, the environment, and human rights broadly" [Org23, 24]. [Org23, 24] While ISO/IEC 23894 offers general AI RM guidance, it follows a traditional, organization-centric approach and uses a risk definition that diverges from the AI Act [KR25, 16]. To address these gaps, a new European standard *JT021024 AI Risk Management* is currently under drafting, with an expected voting date by the end of September 2026 [KR25, 16], emphasizing a product-centric perspective focused on impacts to health, safety, and fundamental rights.

This section first briefly discusses different approaches to AI RM based on a comparative study published by the Organization for Economic Co-operation and Development (OECD). They analyze global RM frameworks for commonalities and differences, towards interoperability [Org23]. Next, we outline risk classification according to AI trustworthiness aiming for a generic risk identification structure that equally corresponds to key areas for European standards, fostering broad information alignment. Therefore, we contrast NIST's approach to AI trustworthiness [Tab23] with the ALTAI [Hig20]. Finally we highlight the role of ethical project planning to consider the human influence, which is necessary for risk mitigation.

3.2.2.1. Approaches & Interoperability

The document published by the OECD contrasts different approaches to AI RM, including the EU AI Act's Article 9 (version published in 2023), ISO/IEC 23894 *AI - Guidance on RM*, and the NIST *AI Risk Management Framework* (RMF) [Org23]. Their interoperability framework is depicted in Figure 3.7, aiming to promote TAI through RM (define, assess, treat, govern) along the AI lifecycle. More precisely, achieving accountability necessitates "[...] to follow these steps at each phase of the AI system lifecycle [...]" [Org23, 16]. Overall, the approaches they compared "[...] are generally aligned with four top-level steps: 'DEFINE', 'ASSESS', and 'TREAT' risks, and 'GOVERN' [RM] processes" [Org23, 10].

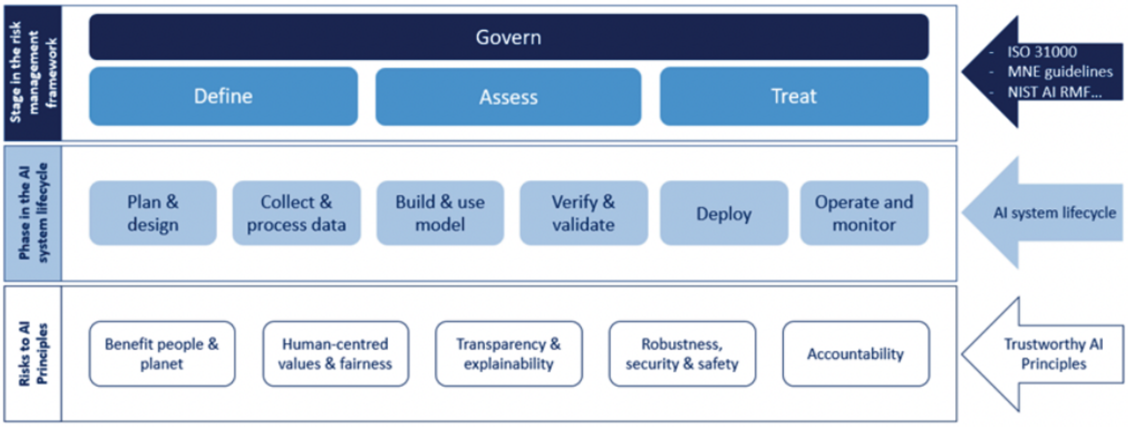


Figure 3.7.: The OECD's "[h]igh-level AI [RM] interoperability framework" [Org23, 19], their depiction.

The first three steps correspond to the previously introduced RM process steps (risk analysis, evaluation, control) and, as introduced, rely on application-specific contextual real world information [Org23, 10]. Most "[...] differences between frameworks relate to the 'GOVERN' function" [Org23, 11], which outlines steps to embed a culture of RM within organizations [Org23, 16]. This includes "[...] monitoring and reviewing the process in an ongoing manner; and documenting, communicating and consulting on the process and its outcomes" [Org23, 16].

- **NIST AI RMF** The NIST AI RMF aligns closely with the DEFINE, ASSESS, and TREAT steps, and both include a GOVERN function. It is designed to manage risks throughout the AI lifecycle, promoting trustworthy and responsible AI systems. Its four core functions (GOVERN, MAP, MEASURE, and MANAGE) address organizational policies, information gathering, risk metrics, and resource allocation. The framework considers impacts on people, organizations, and ecosystems, including human rights and environmental factors. Despite minor differences, e.g. it integrates monitoring, documentation, and communication across all functions (and not only within the GOVERN function) NIST's and OECD's approach share substantial similarities in content. [Org23, 25]

- **ISO/IEC 23894** continues ISO 31000, which aligns well with the DEFINE, ASSESS, and TREAT steps, and partially with GOVERN. Both emphasize embedding RM across organizational levels, stakeholder engagement, and continuous improvement. ISO 31000 focuses more narrowly on organizational risks, sometimes prioritizing organizational value over broader accountability. In contrast, ISO/IEC 23894 broadens the scope to include external stakeholders, societal impacts, and human oversight, aligning more closely with the realization of TAI. [Org23, 23]

- **Article 9 AI Act** The DEFINE, ASSESS, and TREAT steps are included, and Article 9, as previously introduced requires risk analysis, evaluation, and mitigation. Further, parts of GOVERN are addressed through continuous lifecycle monitoring, risk communication, and documentation. [Org23, 27] The AI Act distinguishes itself through its four-level risk classification and the corresponding emphasis on specific requirements for high-risk applications. However, certain GOVERN RM measures, such as stakeholder consultation and integrating RM into organizational culture, appear to be missing [Org23, 11]. In this regard, the AI QMS "[...] comes closest to embedding the [RM] system into broader organisational governance" [Org23, 27].

Overall, both the NIST AI RMF and ISO standards closely align with the OECD's interoperability framework, while the AI Act proposes a slightly different approach but is alignable in combination with the AI QMS on a broad level. Before demonstrating the MQG4AI lifecycle blueprint in section 3.3, which is envisioned to contribute to AI QM through qualitative RAI-knowledge-based information management (IM) for continuous AI lifecycle planning and monitoring, we first concentrate on the initial RM step, i.e. AI risk identification. The next section outlines how to organize potential AI risks to facilitate their integration into specific AI projects, which comprises a fundamental information block of MQG4AI.

3.2.2.2. AI Risk Classification through AI Trustworthiness

Due to AI's evolutionary and opaque character, the technology necessitates continuous RM for individual systems, and concrete risk control measures need to be evaluated towards standardization for (groups of) use cases. Aiming to provide an organized approach to AI RM, and focusing on a comprehensive identification of risks as part of the risk analysis process, we align AI risk classification according to criteria for TAI towards a structured approach for AI RM. From a practical perspective, the most generalizable level of AI-specific risk identification can be addressed through the definition of TAI criteria. Their use case-adapted realization results in negative AI risk reduction. [Tab23, 12] For instance, the NIST AI RMF emphasizes the connection between reducing negative risks and enhancing TAI [Tab23, 12], a concept also reflected by the OECD, as previously introduced and depicted in Figure 3.7, aiming to implement TAI through RM along the AI lifecycle [Org23, 19].

This section briefly contrasts the TAI structure, as introduced by NIST with how the AI Act's ethical foundation is organized. Namely, the "Ethics Guidelines for Trustworthy AI", as defined by an independent high-level expert group (HLEG) [Eur19], which is introduced in sections 3.1.2 and 3.3.5 in more detail. We highlight that all AI risks, including those stemming from AI pitfalls [TJ21] [HSA22] [MF22], for example, can be aligned with the generic and comprehensive HLEG structure.

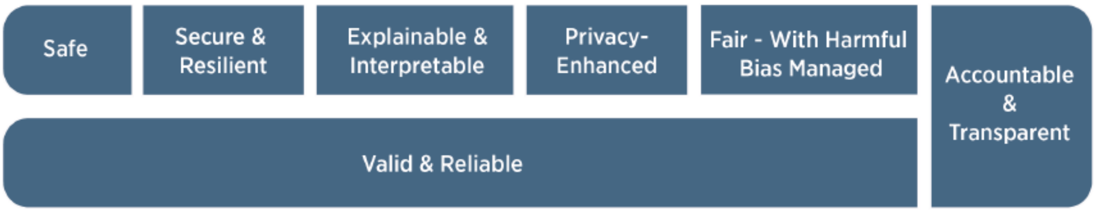

Safe
Secure & Resilient
Explainable & Interpretable
Privacy-Enhanced
Fair - With Harmful Bias Managed
Accountable & Transparent
Valid & Reliable

Figure 3.8.: Key components of trustworthy AI as defined by NIST [Tab23, 12].

TAI for AI Risk Classification The NIST compilation of TAI criteria with related attributes, as depicted in Figure 3.8, is semantically very similar to the AI Act's ethical foundation that comprises seven key requirements (*Human agency and oversight*, *Technical robustness and safety*, *Privacy and data governance*, *Transparency*, *Diversity, non-discrimination and fairness*, *Societal and environmental well-being*, *Accountability*). Both perspectives are differently structured, and we propose that the HLEG's hierarchically structured approach is more applicable to function as a generic and comprehensive risk classification structure. They broach all directions that are related to fundamental rights, and include a compilation of sub-criteria for each TAI criteria [Hig20], which we introduce as foundation for risk classification in section 3.3.5. This hierarchical organization possibly enables the customizable addition of more sub-criteria, answering to AI's evolutionary and dynamic character for risk identification.

Key Differences in Organization As already mentioned, we focus on the overall structure and whether it is applicable to classify evolving AI risks for individual AI projects, and we consider the HLEG's approach to provide a more comprehensive structure compared with NIST. The desired risk classification should enable a holistic overview of all directions for thought that comprise relevant information for AI risks. Within individual projects, they are tailored to the concrete lifecycle implementation, so that design decisions functioning as risk controls can be directly linked. The following list presents key organizational differences.

1. **Valid and reliable outcomes are highlighted as foundation for all TAI criteria**: We agree, valid and reliable results are a global necessity with respect to information flow, interdependencies and stakeholders along the AI lifecycle. Following the HLEG, this requirement is implicitly included with the need for technical and non-technical methods that implement the seven key requirements, focusing on the role of continuous testing and validation [Eur19, 22], which is required for AI RMS processes, as introduced in the previous section.

2. **Accountability is considered a relevant attribute for all TAI criteria**: Semantically speaking, the HLEG equally introduces this seventh requirement *Accountability* as complementing the other requirements [Eur19, 19], but it is listed as stand-alone, which we support for the related risk-collection. In addition to reporting and trade-offs, they highlight "auditability" and an option to "redress", which is made transparent to stakeholders [Eur19, 20] as sub-criteria. These are relevant criteria to enable the implementation of RAI systems [DRN23, 18] through establishing channels that enable taking responsibility for the quality of the intelligent system.

3. **Transparency is considered a relevant attribute for all TAI criteria**: Broadly speaking, according to NIST, "[t]ransparency reflects the extent to which information about an AI system and its outputs is available to individuals interacting with such a system" [Tab23, 15]. This aligns with the HLEG, who broadly address "[...] transparency of elements linked to an AI system: the data, the system and the business model" [Eur19, 18]. The HLEG highlights the traceability requirement as fundamental part of transparency in addition to explainability. Further, they highlight the importance of transparency about the use of an AI system to the human user, and that "the option to decide against this interaction in favour of human interaction should be provided where needed to ensure compliance with fundamental rights" [Eur19, 18]. For risk classification, we propose to treat transparency-related risks as isolated.

4. **Explainability and interpretability are considered separately from transparency**: As equally stated by NIST, both concepts are closely related to transparency: "Transparency can answer the question of "what happened" in the system. Explainability can answer the question of "how" a decision was made in the system. Interpretability can answer the question of "why" a decision was made by the system and its meaning or context to the

user" [Tab23, 17]. The HLEG designates explainability as a sub-criterion of transparency, which we endorse for risk classification along the AI lifecycle, as both concepts align closely in their objectives compared to other TAI criteria.

5. **Security and resilience are mentioned explicitly, and not organized under safety and technical robustness**: "Security and resilience are related but distinct characteristics. While resilience is the ability to return to normal function after an unexpected adverse event, security includes resilience but also encompasses protocols to avoid, protect against, respond to, or recover from attacks" [Tab23, 15]. The HLEG adds this perspective to "Technical Robustness and Safety" [Eur19, 16], which again postulate a similar direction compared with the other six criteria for risk classification. Therefore, collecting relevant risk information in one place is deemed reasonable.

6. **Safety as stand-alone** The safety-attribute requires the AI development process to ensure that the system's foreseeable states do not cause harm and are thus reliable. For instance, reliable performance metrics are crucial to ensure a safe interaction with the intelligent system. Concrete realizations of safety propagating processes are developed within the respective context and the assessed severity of identified risks that could harm a safe and reliable application. [Tab23, 14] In the context of "safety and technical robustness", *accountability and transparency*, as well as *validity and reliability* are explicitly referenced as "accuracy" and "reliability and reproducibility" by the HLEG, in addition to security [Eur19, 17]. They comprise separate subcategories. We support integrating these requirements, as safety and security measures align in a similarly technical direction compared to the other six requirements for lifecycle-adapted risk classification.

7. **Data governance is not explicitly mentioned** Intelligent systems are trained on data sets that can contain sensitive data, especially in the medical context, which results in a number of possible risks regarding "[...] human autonomy, identity, and dignity" [Tab23, 17]. Consequently, implementing means to protect privacy is important to ensure trustworthiness. They "[...] typically address freedom from intrusion, limiting observation, or individuals' agency to consent to disclosure or control of facets of their identities (e.g., body, data, reputation)" [Tab23, 17]. In addition to privacy and data protection, the importance of a comprehensive and trustworthy data governance that addresses data quality, integrity, and access management [Eur19, 17] is emphasized by the HLEG, and highlighting data in relation to privacy is reasonable for risk organization.

8. **Harmful bias is related with fairness** Human-centered values and fairness are of global importance regarding AI, as stated by the OECD with 37

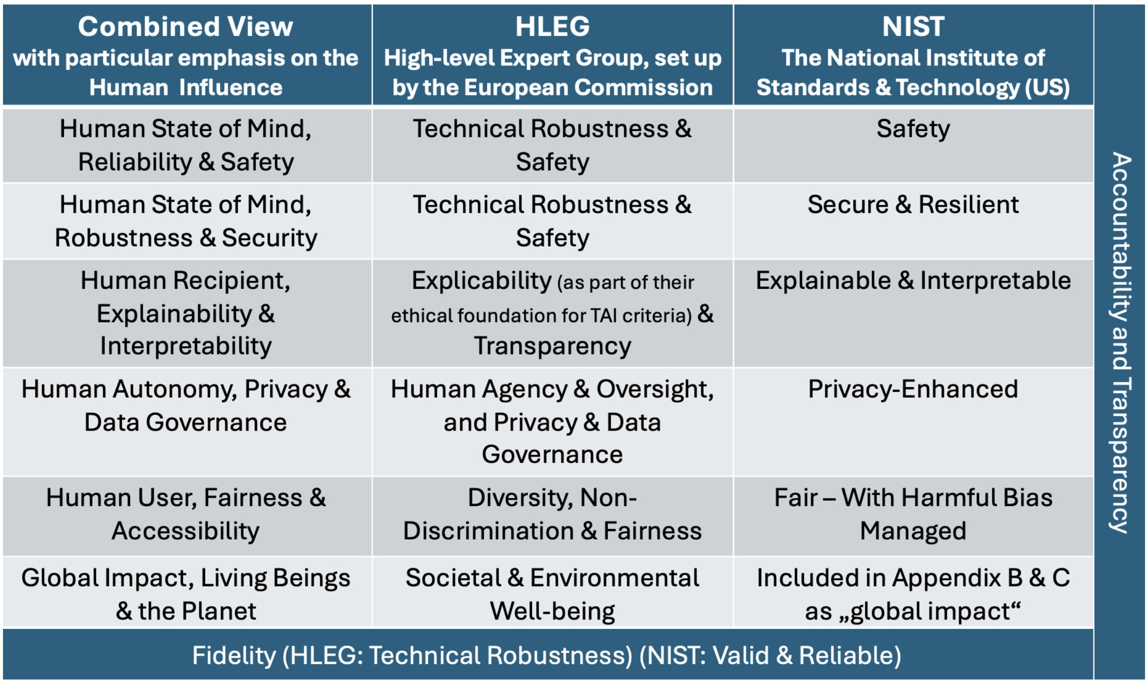
participating democracies worldwide.[10] Especially, for high-risk domains such as medicine, unwanted biased behavior can result in harmful risks for many human beings. Challengingly, the concept of (un)fairness is characterized by societal factors and can change over time and from region to region, which necessitates a dynamic approach. Further, "AI systems can potentially increase the speed and scale of biases and perpetuate and amplify harms to individuals, groups, communities, organizations, and society" [Tab23, 18]. In addition to unfair bias mitigation, the HLEG highlights universal accessibility to "the widest possible range of users" [Eur19, 19], as well as stakeholder participation [Hig20, 18] as crucial components of the broader context of non-discrimination. We agree on incorporating an umbrella term for unfair bias, recognizing that biases may arise from different sources and at various stages of the AI lifecycle, and that their classification is often complex. Refer to [EM25b], where we outline our considerations on how to approach bias during AI projects focusing on the human influence. As a result of this multi-faceted setting, bias-associated risks could be linked as sub-categories assigned to other TAI criteria, following the HLEG's structure for unfair bias. For instance, biases may emerge at earlier lifecycle stages, such as data bias, or later stages, such as automation bias during operation when the human user tends to blindly trust the system's output.

9. **Human agency and oversight is not explicitly mentioned** This is a key difference in both approaches, and NIST does not mention human agency as part of their TAI criteria list. While, they mention that "[s]ome AI systems may not require human oversight, such as models used to improve video compression. Other systems may specifically require human oversight" [Tab23, 40], which is the case for high-risk AI, as defined in the AI Act. The HLEG focuses on the ethical requirement for a reliable real world integration, emphasizing the protection of fundamental rights [Eur19, 15]. Intelligent systems "[...] should support human autonomy and decision-making" [Eur19, 15], and they introduce different formats to implement human oversight [Eur19, 16]. Those requirements necessitate a visible spot within an AI risk classification to link implemented controls along the AI lifecycle, aiming to enable human stakeholders, including the provision of relevant information.

10. **Societal and environmental well-being is not explicitly mentioned** NIST did not explicitly include such requirements in their TAI list but concretized global impact in Appendix B and C as crucial parts of AI risks [Tab23], also they include "people and the planet" as central key dimensions around which the AI lifecycle is built [Tab23, 10]. The HLEG includes this universal perspective as part of their key requirements for TAI. And we concur that social and environmental well-being, which refers to valuing sustainability and addressing the societal impact of the intelligent system [Eur19, 15]

[10] https://oecd.ai/en/ai-principles

deserves a prominent placement within an AI risk classification that corresponds to risk-mitigated design decisions along the AI lifecycle.



Combined View with particular emphasis on the Human Influence	HLEG High-level Expert Group, set up by the European Commission	NIST The National Institute of Standards & Technology (US)	
Human State of Mind, Reliability & Safety	Technical Robustness & Safety	Safety	Accountability and Transparency
Human State of Mind, Robustness & Security	Technical Robustness & Safety	Secure & Resilient	
Human Recipient, Explainability & Interpretability	Explicability (as part of their ethical foundation for TAI criteria) & Transparency	Explainable & Interpretable	
Human Autonomy, Privacy & Data Governance	Human Agency & Oversight, and Privacy & Data Governance	Privacy-Enhanced	
Human User, Fairness & Accessibility	Diversity, Non-Discrimination & Fairness	Fair – With Harmful Bias Managed	
Global Impact, Living Beings & the Planet	Societal & Environmental Well-being	Included in Appendix B & C as „global impact“	
Fidelity (HLEG: Technical Robustness) (NIST: Valid & Reliable)			

Figure 3.9.: A possible combination of the HLEG's [Eur19] and NIST's [Tab23] approach to summarize TAI criteria, with a particular emphasis on the human influence.

Additionally, aligning TAI-related risk identification during intelligent system development with the *Sustainable Development Goals*[11], as imposed by the United Nations (UN), can support risk identification and achieving long-term success. Figure 3.9 illustrates the previously compared TAI perspectives, and highlights a possible combination through the lens of the human influence with respect to each TAI criteria, from the developer's viewpoint. Safety, security, and related concepts, such as resilience, are intrinsically connected to the human state of mind. This connection becomes particularly evident in scenarios like responding to phishing emails or evaluating foreseeable risks and potential harm during RM. Moreover, explainability and interpretability emphasize approaches designed with the human recipient in mind, while privacy and data governance provide a critical foundation for safeguarding human autonomy. Fairness and accessibility, in turn, are closely linked to the diverse user personas interacting with intelligent systems. Overall, humans orchestrate the processes that result in individual, risk-mitigated AI lifecycles and interact with the system. Since the human mind is complex and can be vulnerable under certain conditions, and humans are biased by default, it needs to be considered to implement RAI. Together, these perspectives converge to establish a robust and holistic RMS.

[11] `https://sdgs.un.org/goals`

3.2.2.3. The Human Influence & Ethical AI Project Planning

Given AI's transformative potential, we argue that considering the human influence throughout the AI lifecycle, particularly in high-risk applications like medicine is essential to ensuring trustworthy and compliant AI. Due to AI's inherent dynamics such as its opacity, or stochasticity, implementing the rather abstract TAI concepts is not trivial for the multitude of possible use cases, and concrete design decisions are made by human beings. As a result, the human influence comprises a crucial component of the AI lifecycle. This includes RM processes that are characterized by the respective use case, as outlined in section 3.2.1.4. This section briefly introduces the human influence, and highlights the relevance of ethical project planning. We propose a holistic approach to foster a RAI mindset, and we emphasize the importance of ethics training for contributing stakeholders as a risk mitigating measure as part of continuous AI lifecycle planning. MQG4AI, which is envisioned to provide a comprehensive AI lifecycle planning information structure, including ethics, is introduced in the next section.

Contributing Stakeholders AI actors are defined as "[...] those who play an active role in the AI system lifecycle, including organizations and individuals that deploy or operate AI" [Tab23, 2] in the NIST AI RMF, following OECD. We refer to those as *contributing stakeholders*, such as developers, regulators, or domain experts, which we further explore in section 3.3.3.5. Contributing Stakeholders bear responsibility regarding the impacts of the intelligent system they implement and they contribute information based on their current level of knowledge. This terminology is intended to highlight the TAI-criteria *Accountability* [Eur19, 19], which is equally reflected by the OECD's AI principles [Org23, 16], and summarized in the RAI requirement to implement accountable systems, as demonstrated in section 3.1.3. In addition, "[...] AI [RM] depends on a sense of collective responsibility among [contributing stakeholders]" [Tab23, 10]. For instance, the prioritization of TAI criteria is domain- and use case-dependent and includes the involved human along the lifecycle [Tab23, 12]. Also, concrete design-decision making how to implement lifecycle stages in a risk mitigated manner belongs to contributing stakeholders. Further, human beings are inherently biased, and biases, which pose risks if they result in unwanted behavior are mostly based on human sources [ISO21a], as introduced in detail in [EM25b]. In summary, to complement the implementation of the previously outlined RM processes, this human influence necessitates ethical awareness among contributing stakeholders, highlighting the need for ethics training.

Ethical Project Planning Currently, there is a "[...] lack of comprehensive interdisciplinary approaches that could support teams in integrating ethical considerations into the agile software development process" [KJ24, 1], which is extendable to AI projects. To foster ethical project planning, we are convinced that embedded ethics [MS20] along the AI lifecycle development process is a necessary step towards RAI. Stakeholders at all stages of lifecycle planning and execution are advised to ask and discuss questions related to TAI when defining concrete design decisions from abstract planning to use case-specific implementations. Generally, ethical concepts need to be learned and understood on an abstract level and with respect to the concrete project.

Holistic Approach In [EM25b], we introduce a holistic approach towards RAI that addresses the human influence on the AI system and its ethical implications, aligning with the EU AI Act's risk-based framework. Closing the gap from an ethical foundation to an applicable ethos, materials for ethics training are appended, aiming to encourage interdisciplinary collaboration and an adaptable, risk-aware AI development process that reflects both technical and societal values. We outline four key pillars foundational to RAI: *Generalizability*, *Adaptability*, *Transversality*, and *Translationality*, as depicted in Figure 3.10. These related pillars provide a structured approach, envisioned to guide stakeholders towards a dynamic, ethical, and compliant AI lifecycle that aligns with regulatory standards and societal expectations. The proposed approach is underpinned by continuous *Education* and *Research*, fostering a "RAI mindset" that prioritizes ethics in AI project planning.

***Transversality* to address the multi-faceted concept of bias** A key innovation of this work is the introduction of *Transversality* as a foundational pillar. The concept can be traced back to the post-modern era, and was adapted from application in media philosophy to AI ethics. *Transversality* offers a novel terminology to describe the multi-faceted concept of bias in its completeness by embracing dynamic reasoning across contexts, acknowledging that biases are not merely technical issues but are deeply influenced by societal values and human cognition. Unlike "bias" or "fairness" that may compartmentalize or overlook some facets of bias, *Transversality* encourages stakeholders to consider a pluralistic perspective, adapting to AI's role within diverse social and ethical contexts.

The other three pillars *Generalizability* ensures AI models can perform consistently across various conditions, supporting their reliability. *Adaptability* emphasizes the need for models to evolve as new data and real world contexts change, addressing AI's inherent dynamics. *Translationality* bridges theoretical AI concepts to practical applications, ensuring seamless integration with real world systems.

Embedded Ethics Aiming to close the gap to applied ethics during RAI project planning and execution, we illustrate a possible application scenario based on the DARE-method [KJ24] that considers embedding ethics with agile software development. In agile development, the *product owner* serves as the central interface between the product team and various stakeholders, including consulting, passive, external, or other active participants involved in the project. A key respon-

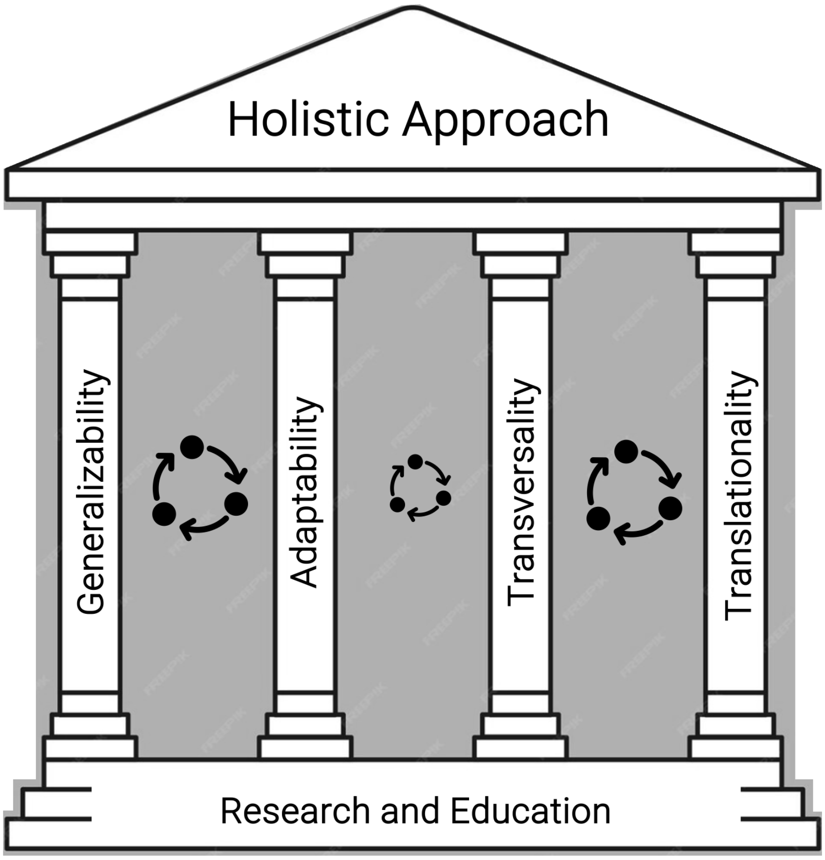


Figure 3.10.: The proposed holistic approach to foster a RAI mindset for developing intelligent systems that provide stability and flexibility regarding their real world behavior, while ensuring seamless integration into real world contexts from both technical and social perspectives. The four interconnected pillars require ongoing evaluation and alignment with the AI system's intended environment of use.

sibility of the *product owner* is to effectively convey the project vision, ensuring alignment and shared understanding across all parties. We propose integrating knowledge of RAI into this role, or the inclusion of an extra RAI person. Figure 3.11 illustrates the agile development process and highlights stages where ethical considerations within the team become particularly relevant:

- (A) Focuses on the "what" of development, addressing goals, requirements, and the overall product direction.

- (B) Emphasizes the "how," centering on processes, methods, and approaches to achieve these goals.

- (C) Represents a reflective stage before the next iteration, aimed at evaluating and adapting based on ethical and practical insights.

This structured approach integrates ethical considerations into the agile workflow, fostering responsible and thoughtful development practices. DARE, as illustrated in Figure 3.12, emphasizes key topics such as training data quality and setup, the role of the human-in-the-loop, and explainability as core pillars of RAI. These elements act as foundational anchors for addressing the ethical, technical, and practical challenges inherent in AI development. By centering on these abstract yet critical concepts, DARE establishes a dynamic framework for RAI discussions that evolves in tandem with a project's iterative cycles. Combined with specific flashcards, this framework provides a solid foundation for translating RAI principles into practical applications tailored to individual projects. Ultimately, this approach encourages thoughtful dialogue and supports the seamless integration of RAI values across diverse development scenarios.

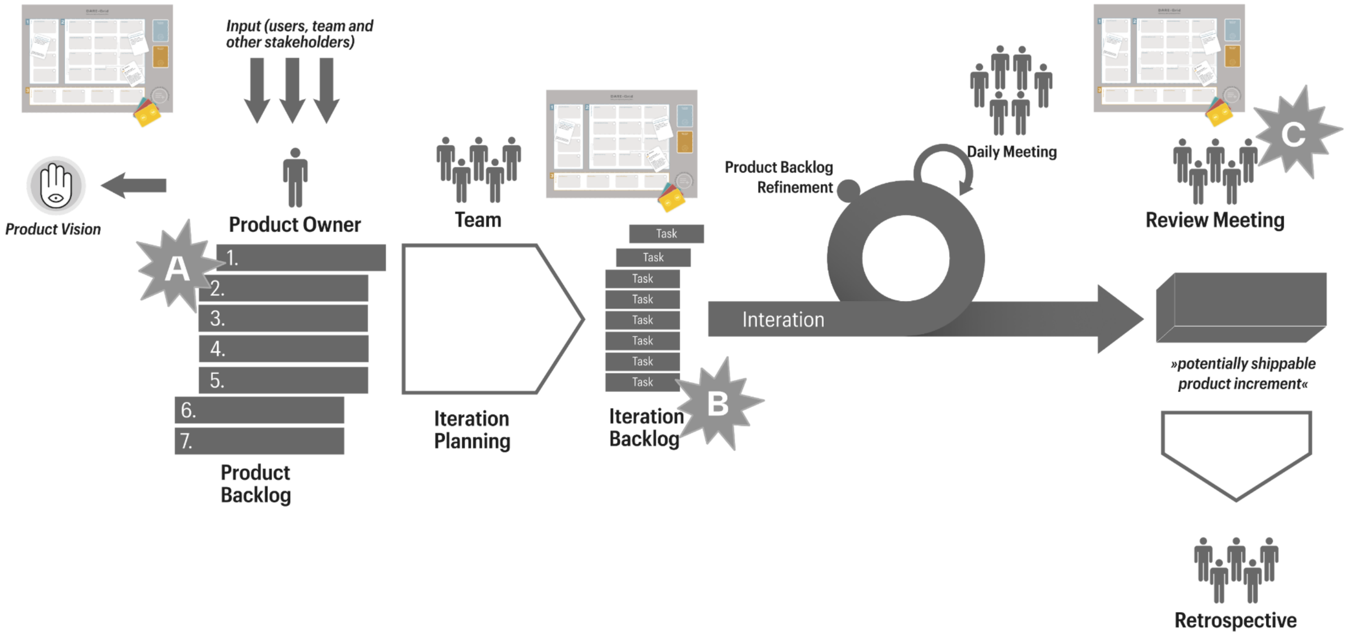


Figure 3.11.: Ethics training based on DARE [KJ24] integrated with agile product development.

Considering the Human User Finally, regarding the system's integration into real world applications, the human user needs to be accounted for early from the start of AI project planning by contributing stakeholders, when designing the lifecycle including methods to implement TAI. Regarding a safe interaction, for instance, it is important to consider the state of the system's human user [Eur19, 16] as an additional perspective towards comprehensive RM. NIST mentions the importance of a "safety-first mind-set" [Tab23, 23] as part of their overall framework, and they emphasize the importance of a clearly defined transfer of information to the human user. Risks can emerge based on an inadequate understanding of the system's inner workings or an inability to correctly interpret its outcomes.

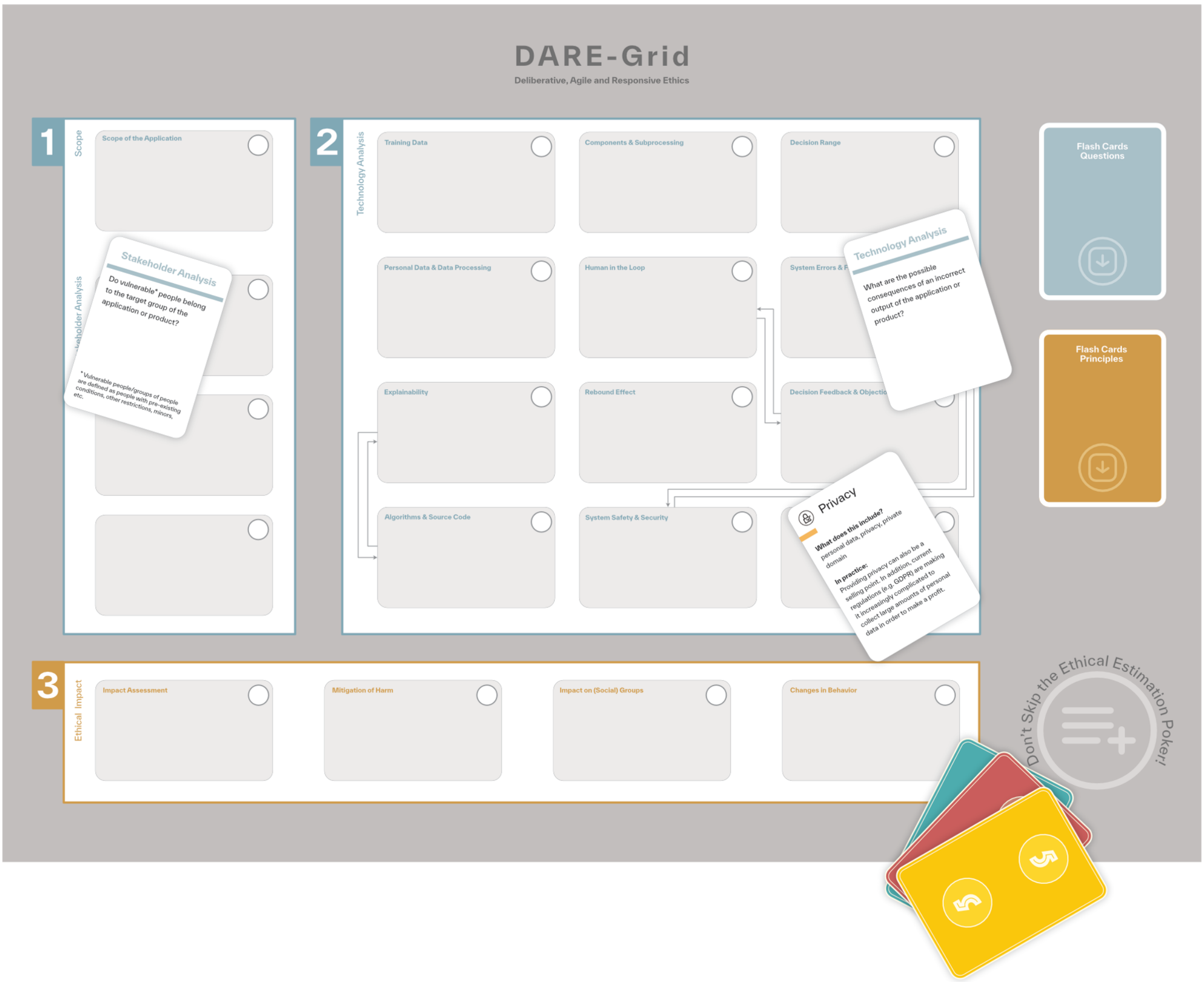

DARE-Grid
Deliberative, Agile and Responsive Ethics
1
Scope
Scope of the Application
Stakeholder Analysis
Stakeholder Analysis
Do vulnerable* people belong to the target group of the application or product?
*Vulnerable people/groups of people are defined as people with pre-existing conditions, other restrictions, minors, etc.
2
Technology Analysis
Training Data
Components & Subprocessing
Decision Range
Personal Data & Data Processing
Human in the Loop
System Errors & F
Technology Analysis
What are the possible consequences of an incorrect output of the application or product?
Explainability
Rebound Effect
Decision Feedback & Objectio
Algorithms & Source Code
System Safety & Security
Privacy
What does this include? personal data, privacy, private domain
In practice: Providing privacy can also be a selling point. In addition, current regulations (e.g. GDPR) are making it increasingly complicated to collect large amounts of personal data in order to make a profit.
Flash Cards Questions
Flash Cards Principles
3
Ethical Impact
Impact Assessment
Mitigation of Harm
Impact on (Social) Groups
Changes in Behavior
Don't Skip the Ethical Estimation Poker!

Figure 3.12.: The DARE-grid, and flashcards, illustrating the *gamification*-approach towards embedded ethics [MS20].

Enhancing usability, by designing systems that are more transparent and interpretable, can serve as an effective risk control measure. However, achieving this is far from trivial. For example, if a training dataset is predominantly composed of data from white patients, it is essential to inform medical professionals that the system's predictions may be less reliable for patients of color. Clear communication of such limitations, complementing the AI system helps users make informed decisions and mitigates potential risks.

In summary, this section highlights the human influence and relevance of ethical project planning, and we propose a holistic approach that can function as foundation to consolidate TAI concepts for contributing stakeholders during individual projects, aiming for on-going awareness. After detailing RAI and related concepts, with a particular emphasis on RM and TAI, the next section introduces MQG4AI. The lifecycle planning blueprint is based on a comprehensive information structure that considers technical interdependencies, as well as contextual real world information to support high-quality AI lifecycle implementations.

3.3. MQG4AI – Towards Comprehensive & Continuous AI Lifecycle Design

This section introduces the basic information structure that serves as the foundation for the lifecycle planning blueprint of our proposed generic and customizable methodology based on Quality Gates (MQG4AI). We envision to contribute towards implementing responsible high-risk AI, as previously introduced in section 3.1, with a particular emphasis on RM, as outlined in section 3.2. We focus on certifiable AI in medicine through the use cases introduced in section 5 and the evaluation of MQG4AI interaction scenarios, as outlined in part III, while the concept design aims for generalizability across domains.

Overall, we envision to provide an information management (IM) structure that contributes to AI Act-conform Quality Management Systems (AI QMS), as clarified in Article 17 [Fut24]. Therefore, this work adopts a development-centered perspective situated at the interface with regulatory compliance. The proposed approach is grounded in comprehensive lifecycle planning and RAI knowledge management (KM), which also promotes AI literacy in alignment with the requirements set out in Article 4 of the AI Act [Fut24]. The generic structure of the AI lifecycle, whose robust design constitutes the core of the MQG4AI blueprint, is illustrated in Figure 3.13.

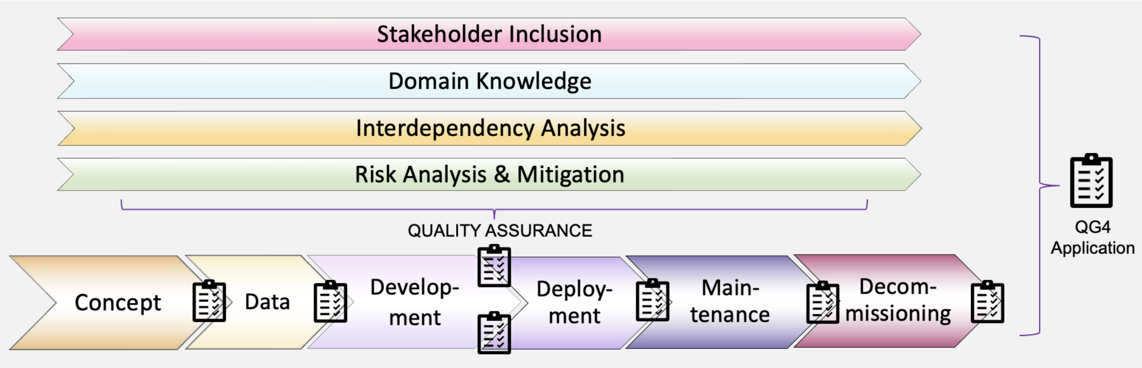


Figure 3.13.: Identified generic AI lifecycle stages, answering the questions *Where is the model and what does it need?* Different phases and design decisions are mirrored by Quality Gates (QG) that comprise the foundation to design the MQG4AI blueprint. The overall methodological design is guided by four fundamental design principles, as introduced in section 3.3.2.

Development We concentrate on *AI Provider* as a result of our implementation-centered lifecycle view, with a particular focus on the underlying design concepts and their inter-relations regarding outcome quality. *AI Provider*[12] are defined as "[...] a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge;" in Article 3 of the AI Act [Fut24]. It is the provider's responsibility to prove the required quality, and, among other requirements, AI QMS include a risk management system (RMS) according to Article 9 [Fut24]. Contributing information to RM comprises a primary objective of MQG4AI design.

Regulation In addition, MQG4AI is intended to provide relevant information to additional stakeholders during *conformity assessment*, i.e. "the process of demonstrating whether the requirements set out in Chapter III, Section 2 relating to a high-risk AI system have been fulfilled", such as a *Notified Body*, i.e. "a conformity assessment body notified in accordance with [the EU AI Act] and other relevant Union harmonisation legislation". A *conformity assessment body* "means a body that performs third-party conformity assessment activities, including testing, certification and inspection", as defined in Article 3 of the AI Act. [Fut24]

We first define our interpretation of IM, which follows ISO/IEC FDIS 5338:2023 on "AI System Lifecycle Processes" [ISO23c], and how this relates to RAI KM [ISO23c]. Next, we highlight four fundamental design principles for MQG4AI design towards overall risk mitigation, as depicted in Figure 3.13, and demonstrate customizable RAI information blocks. The current version of the MQG4AI blueprint consists of three information blocks, which are introduced in more detail in the following sections, namely *System*, *Risk Management*, and *Lifecycle* information. Other IM modules are envisioned to be appended analogously, focusing on information linking with lifecycle design. Next, we introduce our methodological foundation *Design Science Research* [vJ20], and the proposed MQG4AI interaction scenarios towards continuous and decentralized blueprint design. Finally, we outline our vision for and limitations of MQG4AI in section 3.4, before diving deeper into its conceptual setup and building blocks in section 4.1.

3.3.1. Evolving RAI Knowledge through Information Management

The notions *information* and *knowledge* are very similar, and "[...] there is no single, general definition or distinction to be made. There rather seems to be various differences together with similarities between the two notions" [Bra02, 80]. A key difference lies in the human factor, and *information* can "[...] exist independent of

[12] Or, *AI Developer*, as defined in ISO 22989 [ISO22a].

a subjective holder" [Bra02, 88]. Consequently, *knowledge* is derived from *information*, which is based on facts that can be extracted from data, documents, or other resources [ISO22a, 18]. In the context of AI systems, *knowledge* is defined as "[...] abstracted information about objects, events, concepts or rules, their relationships and properties, organized for goal-oriented systematic use" [ISO22a, 4]. This definition is transferrable to human knowledge on AI lifecycle implementations in that it is technical and directed in nature, as well.

MQG4AI is centered around the assumption that an AI system equals the combination of its underlying lifecycle design decisions. Therefore, the AI lifecycle is documented in a structured manner to produce *information* that fills the MQG4AI blueprint in an evolving manner, while adhering to the technology's inherent dynamics. Further *knowledge* to develop/optimize/monitor the AI system, as well as to conduct quality assessment is derived from the thus generated *information* on the concrete implementation of the individual project. Depending on the respective stakeholder, this *information* is represented differently to produce *knowledge*. The proposed structure in form of MQG4AI's building blocks, as outlined in chapter 4, provides *information processing layers* for design decision-making along the AI lifecycle. In addition, different template versions, that are pulled from the living blueprint and introduced in section 3.3.7, offer means to incorporate knowledge on lifecycle evolutions through organizing information, aiming for a continuous conceptualization phase.

This section first introduces IM and how MQG4AI relates with RAI KM processes, attempting to convey the meaning of *knowledge* and *information* in the context of MQG4AI. Overall, the blueprint is envisioned to provide a lifecycle planning structure that enables AI IM processes focusing on the intersection between AI providers and regulation from the implementation-perspective through the dynamic incorporation and generation of *RAI Design Knowledge*. Therefore, we outline RAI knowledge contributions to design MQG4AI next, before introducing MQG4AI's fundamental design principles in the following section.

3.3.1.1. Information & Knowledge Management

Our proposition continually refines knowledge on the AI system, grounded in evolving design KM and organized IM. Therefore, we emphasize the importance of documenting the complex and dynamic lifecycle concept, as this process translates human expertise in AI design into structured information, which subsequently cultivates stakeholder knowledge about specific implementations. Moreover, integrating lifecycle information with contextual details and addressing technical interdependencies contributes to improving the overall quality of the system.

In contrast to our lifecycle perspective centered around contributing stakeholders, ISO 22989 defines *Knowledge* with respect to the AI system: "Knowledge differs from information in that information is observed by the system, while knowledge is what the system retains from such observations" [ISO22a, 18]. This abstract definition is transferable to human knowledge on AI design, which shapes/is retained from information on the AI system implementation. Further, "[t]he model used by the AI system for its processing and for problem-solving is a machine-readable representation of knowledge" [ISO22a, 41]. In the context of DNNs, translating the AI system's knowledge to humans is not trivial due to the model's inherent opacity. The model is built upon the design knowledge contributed by stakeholders involved in shaping the AI lifecycle, alongside information derived from the underlying data and use case, which is equally prepared by contributing stakeholders. Therefore, the MQG4AI lifecycle planning blueprint is interpreted as a human-readable representation of knowledge on the intelligent system's implementation, embedded within contextual information, while updating and profiting from globally existing RAI design knowledge. This organized information, and the derived human knowledge on the system's design are envisioned to contribute to shedding light on the model's representation of knowledge, enhancing overall transparency to support quality assessment and risk mitigation. This is achieved through IM and KM, and the cyclic interplay between RAI design knowledge and information from the developer's perspective is illustrated in Figure 3.14.

Information Management In the Cambridge Dictionary, IM is defined as executing the organization of existing information, it is "the process of collecting, organizing, storing, and providing information within a company or organization".[13] Tailoring this definition to the AI lifecycle, ISO 5338 on *AI system lifecycle processes* defines the lifecycle IM process as follows, highlighting the respective information's target group:

> The purpose of the [IM] process is to generate, obtain, confirm, transform, retain, retrieve, disseminate, and dispose of information, to designated stakeholders. The [IM] process plans, executes, and controls the provision of unambiguous, complete, verifiable, consistent, modifiable, traceable, and presentable information to designated stakeholders. Information includes technical, project, organizational, agreement, and user information. Information is often derived from data records of the organization, system, process, or project. [ISO23c, 16]

In case of MQG4AI, IM comprises the reliable organization of knowledge on how the AI system lifecycle is planned and implemented in relation to supplementary contextual information, such as AI risks. MQG4AI is intended to organize implementation-relevant information for contributing stakeholders that play an

[13] https://dictionary.cambridge.org/dictionary/english/information-management

active role along the AI lifecycle. Consequently, we refer to blueprint components as *information blocks*, and IM describes MQG4AI's functionality.

Addressing AI's inherent dynamics, the blueprint's structure and its customizable building blocks, as introduced in this chapter and the next chapter 4, are intended to enable AI IM that adheres to RAI knowledge towards quality by design based on comprehensive information processing and linking. For instance, IM is achieved through the *leaf-QG format* that mirrors concrete design decisions and processes information on their definition, including the documentation of relevant input and output information for design-decision making. In addition, *collection-QGs* construct the lifecycle in an adaptable manner to allow for the identification of interdependencies. Further, MQG4AI's *Design Science Research* (DSR) [vJ20] setup (MQG4AI interaction scenarios, see section 3.3.7), is intended to adhere to the technology's use case-specificity and evolutionary character that shape lifecycle design choices in an on-going manner. In summary, RAI lifecycle design is based on the evolving knowledge of contributing stakeholders that actively partake in implementing the AI lifecycle. In addition, conformity assessment bodies must have adequate RAI design knowledge to verify that AI systems meet the AI Act's standards for fundamental rights, safety, and health [Fut24].

Knowledge We highlight the distinction between *tacit* and *explicit*, or *implicit* knowledge, and a comprehensive view on types of knowledge is explained in [Nic00]. The latter are or can be articulated, while the former is challenging to transform into tangible information [Nic00, 3]. For instance, DNNs that process data can be described as an "[...] implicit encoding of knowledge" [ISO22a, 20], while explicit encoding refers to formal representations [ISO22a, 20]. Tacit knowledge is not referred to in the context of the AI system, since it relates to how humans perceive the world, i.e. "[...] situations in which we are able to perform well but unable to articulate exactly what we know or how we put it into practice" [Nic00, 3]. In the context of MQG4AI and RAI Design Knowledge for AI lifecycle planning, *explicit knowledge* is summarized by existing publications, analyzable data, and interviewing experts [ISO23c, 21]. *Implicit knowledge* reflects empirical experience, "[...] studying, interviewing or other knowledge elicitation, data analysis, acquiring documented knowledge or involving stakeholders with the required knowledge" [ISO23c, 21], which can be transformed into explicit knowledge. *Tacit knowledge* comprises the individual stakeholder's approach towards conceptualizing the respective lifecycle stage, and design decision. All three types comprise relevant RAI knowledge that is (partly) transformed into information on a reliable lifecycle implementation through different interaction scenarios with the MQG4AI blueprint. Especially, *tacit knowledge* "[...] is closely bound to the person who developed it; it is shared primarily through person-to-person contact" [Org04, 20], and the MQG4AI-blueprint is envisioned to capture parts of the individual's approach to lifecycle design decision-making, based on a flexible and dynamic RAI information structure that enables use case-specificity, while providing general guidelines.

Knowledge Management RAI knowledge comprises the foundation to fill the MQG4AI blueprint with information, which in turn generates knowledge on the individual system for different stakeholders, as well as may contribute global design knowledge in form of publications, for instance. Consequently, for MQG4AI to be successful, the human factor plays a crucial role, and the MQG4AI blueprint is intended to incorporate *RAI Knowledge Management* through IM. When applying MQG4AI, knowledge on the implementation of individual projects is gathered based on transforming existing RAI knowledge through IM into tangible and measurable formats, which may include data sheets, and AI outputs to create and derive project-specific, possibly novel RAI knowledge. This cycle is depicted in Figure 3.14.

According to the Cambridge Dictionary, KM is defined as "the way in which knowledge is organized and used within a company, or the study of how to effectively organize and use it".[14] This aligns with KM, as defined by ISO 5338, which describes the overall structure on how to approach the organization of RAI knowledge and related information in a sustainable manner. In light of AI's character, promoting AI literacy through dynamic RAI knowledge integration comprises an objective of MQG4AI design:

> The purpose of the [KM] process is to create the capabilities and assets that enable the organization to exploit opportunities to reapply existing knowledge. This encompasses knowledge, skills, and knowledge assets, including system elements. [ISO23c, 11]

Different descriptions of KM sub-processes are possible, highlighting knowledge capturing or acquisition, storage, application, creation, and sharing [Org04] [AS19] [AEM18]. We adapt the following summary to MQG4AI, which is based on [AEM18], where the authors conduct a comprehensive literature review to analyze the relation between KM process and their impact on the adoption of information systems. The extracted KM processes include *knowledge sharing*, *knowledge acquisition*, *knowledge application*, *knowledge storage*, *knowledge protection* and *knowledge creation* [AEM18, 177]. For instance, some KM processes are better researched than others, and knowledge sharing, acquisition, and application positively impact the "adoption, acceptance, and implementation" [AEM18, 182] of different information systems [AEM18, 182]. This finding is probably transferable to AI, emphasizing the importance of KM, which plays a significant role regarding innovation management [Org04, 109].

Figure 3.14 depicts KM processes in the context of MQG4AI. Overall, the lifecycle planning blueprint contributes to KM through documenting RAI design knowledge for different stakeholders, and thus producing information on the AI lifecycle from a comprehensive implementation perspective that considers possibly scattered contextual information, aiming towards quality by design.

[14] https://dictionary.cambridge.org/dictionary/english/knowledge-management

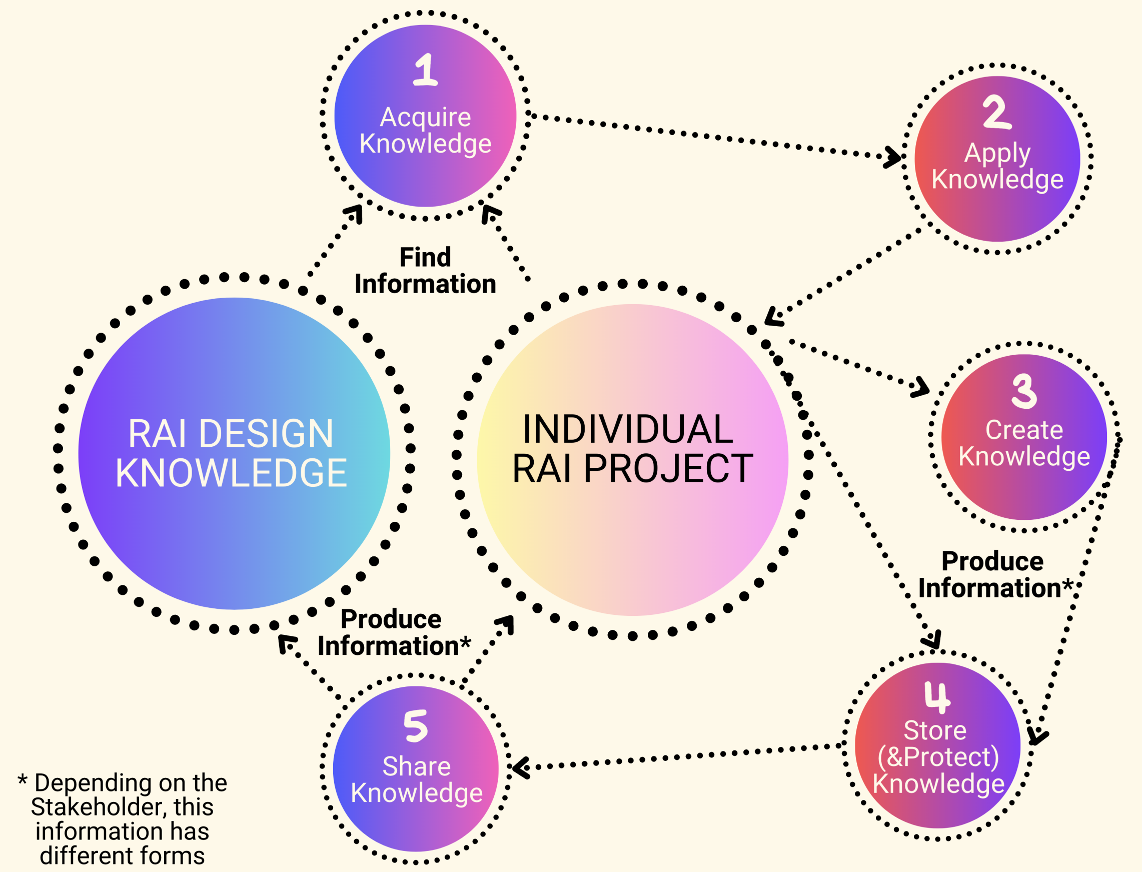


Figure 3.14.: KM processes [AEM18, 177] in the context of MQG4AI from the developer's view on RAI knowledge and lifecycle design decision-making: The developer acquires knowledge on lifecycle design tailored to the respective project through an analysis of existing information [ISO23c, 21], this knowledge is then applied, and lifecycle design decisions are implemented. During lifecycle execution, knowledge is continuously created in the context of the individual project, which is then documented and stored in form of MQG4AI for lifecycle planning, including the applied knowledge. Concurrently, the blueprint provides options for knowledge sharing within the individual AI project, and externally to promote general AI literacy. In summary, this iterative cycle between information and knowledge aims to refine lifecycle design towards achieving the desired quality of the AI system.

1. *Knowledge acquisition* covers utilizing existing and capturing novel knowledge [AEM18, 173], and "[...] all organisations are increasingly dependent on external sources of knowledge" [Org04, 44]. In the context of MQG4AI, *knowledge acquisition* aims to "[...] provide the knowledge necessary to create [and evaluate] the AI models" [ISO23c, 21]. Relevant knowledge is derived from global RAI design knowledge, as well as information on the individual

project that shapes applicable design knowledge.[15] This process involves organizing and integrating relevant knowledge to ensure it becomes a part of the collective knowledge base [AS19, 355] of all contributing stakeholders involved in the individual AI project. Aiming for traceability, outcomes include the identification of relevant RAI knowledge for the use case at hand, as well as *knowledge storing* [ISO23c, 21]. The latter is depicted as an individual process in Figure 3.14, equally capturing *knowledge creation* through *knowledge application*. As a result of the combined KM processes, with growing MQG4AI-application that captures knowledge, further knowledge can be acquired based on previous template versions within an organization, and other template branches within the same project, for instance, in addition to globally accessible knowledge that shapes the living MQG4AI lifecycle blueprint.

2. *Knowledge application* describes processes that make knowledge accessible to stakeholders [AEM18, 173] for *knowledge acquisition*, which enables implementing that knowledge for utilization, resulting in value creation [AS19, 356]. Further, a stakeholder's ability to leverage a relevant knowledge base in decision-making and problem-solving empowers projects to adapt more efficiently to changing environmental factors [AS19, 356]. This evolutionary character is equally relevant to AI from a technical viewpoint, highlighting the connection between *knowledge application* and *knowledge creation* with respect to AI. Finally, marketing RAI *knowledge application* of contributing stakeholders can enhance trust in the approved AI system, analogously to the service industry, who "[...] depend to a great extent upon marketing the application of the knowledge of their workers" [Org04, 56].

3. *Knowledge creation* is especially relevant to AI due to the technology's inherent dynamics, and individual projects necessitate research and testing of methods regarding their applicability to a particular use case and data. For instance, within individual AI projects, different MQG4AI template versions that are based on the continuously evolving blueprint, link outputs with implemented lifecycle states and provide knowledge to different stakeholders. They support learning for AI developers and enable quality insights for conformity assessment, potentially generating novel knowledge for both developers and notified bodies. Overall, *knowledge creation* emphasizes the need for skilled personnel, since it is based on the "[...] capability to generate new applications from existing knowledge and to exploit the unexplored potential of new skills" [AS19, 355]. However, *knowledge creation* appears to be overlooked by professionals [Org04, 34], and so far, is researched less frequently with respect to non-intelligent information systems [AEM18, 183].

[15] For instance, domain-embedded tuning objectives influence the choice of performance evaluation metrics [EM24], and medical domain knowledge on the Hounsfield Unit can optimize preprocessing of CT-scans [MD23].

4. *Knowledge storing* Long-term, *knowledge storing* plays a significant role to ensure the cyclic interplay with all other processes, as depicted in Figure 3.14, aiming towards successful innovation [Org04, 110] through converting knowledge into a format that can be stored, allowing easy access and use by anyone in the organization [Org04, 20]. From a conceptual view, the blueprint provides a structure to store acquired, applied, and created knowledge through RAI IM, which needs to be executed by motivated stakeholders [AS19, 355]. Codified knowledge, or structured information can be stored and reused, but not all knowledge is easily captured or codified. Organizations possess significant implicit, and tacit, or experience-based knowledge that are harder to capture, highlighting the importance of managing both explicit and tacit forms of knowledge. [Org04, 38] MQG4AI is envisioned to support this process through its information structure and, as a result support transparency and enhance AI literary. *Knowledge storing* relates to *knowledge protection*, an equally less well researched process [AEM18, 183], which we do not investigate further, since, e.g. lifecycle information access management is beyond our current scope. Generally, the application of MQG4AI templates is private within an organization, and external knowledge sharing happens in a controlled manner.

5. *Knowledge sharing*, meaning the provision of relevant knowledge to all stakeholders that participate in a process, is at the heart of KM [AEM18, 173]. Generally, *knowledge sharing* has different realizations depending on the type of knowledge shared, for instance, they include "[...] communicating, learning, reviewing, capturing and sharing knowledge" [Org04, 34]. In particular, sharing the tacit knowledge of stakeholders is challenging [AS19, 356]. Therefore, MQG4AI provides accessibility to all stakeholders within a project, analogously to the Git-branching structure[16] on an abstract level, documenting combinations of possible implementation approaches. In addition, since AI is a dynamic and intricate technology, MQG4AI aims to facilitate public knowledge sharing with the broader scientific community, as well as the industry on RAI implementation approaches, incorporating DSR [vJ20].

In summary, MQG4AI is anticipated as an interactive and evolutionary information collection of a qualitative RAI lifecycle design that incorporates and contributes to RAI knowledge for actively involved stakeholders within an organization and globally. IM describes the blueprint's functionality, filled with design knowledge through human interaction, resulting in continuous KM, and ultimately supporting stakeholders in interpreting project-specific information through a unified structure. Following the introduction of MQG4AI's roots in IM and KM, we exemplify RAI knowledge through different resources that together contribute to the design of MQG4AI.

[16] https://git-scm.com/book/en/v2/Git-Branching-Branches-in-a-Nutshell

3.3.1.2. RAI Knowledge for MQG4AI Design

Highlighting the technical nature of knowledge in the context of an AI system [ISO22a, 18], AI lifecycle design knowledge is technical at its core, as well. Multiple sources for AI design knowledge acquisition by contributing stakeholders exist, and examples comprise technical publications by the research community, the industry, as well as (international) standardization bodies for explicit knowledge, or documented information. In addition, empirical AI design knowledge is based on experiences with AI system creation, and assessment, resulting in implicit knowledge, which can be transformed into tangible information. This may include tacit knowledge of an individual, which is harder to articulate.

Aiming for a RAI system implementation, technical design knowledge contributions are aligned with the MQG4AI information structure. It links design decision-making with supplementary information, closing the gap to TAI and the real world through the integration of AI risks and domain embedding, while addressing interdependencies, as starting point towards AI QM. The four fundamental design principles that shape MQG4AI's information structure towards RAI by design are further outlined in the next section 3.3.2. Consequently, RAI Design Knowledge is a combination of existing AI design information, aligned in accordance with MQG4AI's generic and customizable building blocks, aiming for RAI by design. Further, we highlight alignment with existing standards for blueprint design, since we aim to contribute to overall AI QM. However, as introduced in section 3.1.3.2, the current standardization landscape surrounding AI within the EU is in its infancy. Therefore, we intend to provide a flexible lifecycle planning blueprint focusing on IM, that is adaptable to the ever evolving RAI knowledge surrounding this dynamic technology. This involves the ongoing integration and application of information from specific use cases.

MQG4AI	Design Knowledge Source
System Information	The proposed structure of the AI system information block, as outlined in section 3.3.3, is based on the AI Risk Ontology (AIRO) [GD22], which aligns with information required by the EU AI Act [Fut24], highlighting necessary documentation, for instance.
Risk Management	The proposed blueprint organizes TAI criteria according to the *Assessment List for Trustworthy AI* [Hig20], as promoted by the European Commission. Further, MQG4AI is envisioned to enable basic RM processes (analysis, evaluation, control), as introduced in section 3.2.1.

Table 3.1.: A brief overview of design knowledge that contributes to the two illustrated MQG4AI contextual information blocks system and RM information.

This section provides an overview of central RAI knowledge that contributes to MQG4AI design, which globally follows the principles of DSR [vJ20]. Table 3.1 exemplifies the organization of contextual AI system, and RM information, while these blocks are intended to be exchangeable with existing management structures. Table 3.2 comprises knowledge contributions to the central AI lifecycle block, including QG design. The MQG4AI information blocks, and setup are further outlined in the following sections.

MQG4AI	Design Knowledge Source
Lifecycle	The generic AI lifecycle perspective of MQG4AI aims for an implementation view on applicable concepts. It is based on ISO/IEC 5338:2023 on *AI system life cycle processes* [ISO23c] and inspired by [DA22]. Further, this generic high-level structure globally aligns with the OECD's lifecycle [Org23, 16], reflecting ISO/IEC 22989 *Artificial intelligence concepts and terminology* [ISO22a, 36], on which ISO/IEC 42001 on *Artificial intelligence management system* is based [ISO23b, 31], while a few differences emerge based on MQG4AI's lifecycle planning functionality. The lifecycle is constructed with collection-, and leaf-QGs which are more throroughly outlined in section 4.2.
Data	The data stage collection-QGs are based on the *IEEE Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence* [IEE22], while data leaf-QGs are not our focus, and only broached in relation to development.
Development	These development stage collection-QGs are mainly derived from empirical experience [EM24] [MD23] [SF23].
Explanation	We highlight collection-QGs that construct the explanation stage. For this stage, we validate empirical expert knowledge [LGC24] [EM25a] against an existing standard, as explored in more detail in chapter 6. As a result, the proposed process structure is alignable with the *IEEE Guide for an Architectural Framework for Explainable Artificial Intelligence* [Art24].
Leaf-QG	The proposed information structure that reflects lifecycle design decisions, as detailed in section 4.3, aims to implement MQG4AI's fundamental design principles. It is envisioned to provide flexibility to extract information for AI QM, as outlined in Article 17 of the EU AI Act [Fut24]. For instance, this includes post-market monitoring information extraction, linking with risks and risk controls, and information transformation to different stakeholder views. Finally, the leaf-QG-template includes information derived from ML design pattern [WH22].

Table 3.2.: A brief overview of design knowledge that contributes to MQG4AI design of the central AI lifecycle information block.

After clarifying RAI knowledge that contributes to MQG4AI design, highlighting the interface with existing guidelines and standards, the next section introduces four fundamental MQG4AI design principles that are intended to level the path for a continuously evolving RAI information structure that embeds risk mitigation by design.

3.3.2. MQG4AI's Four Fundamental Design Principles

This section introduces MQG4AI's four fundamental design principles that shape the structure of the proposed RAI information blocks, that are introduced in the next sections, as well as the design of individual MQG4AI components, as further outlined in chapter 4. MQG4AI's design philosophy aims towards overall *Risk Mitigation*, which is subdivided into *Stakeholder Inclusion*, and *Domain Embedding* focusing on healthcare, as well as a more technical perspective towards a lifecycle-specific *Interdependency Analysis*.

3.3.2.1. Stakeholder Inclusion

Stakeholder Inclusion is a fundamental principle of MQG4AI design, based on the need for diverse teams, as well as a comprehensive stakeholder analysis in alignment with the individual AI project. The human influence shapes individual design decisions, as outlined in section 3.2.2.3 and stakeholders contribute important information to AI lifecycle design.

Stakeholder Analysis Taking all stakeholder perspectives into account is crucial to ensure that the intelligent system fulfills its intended purpose in its (clinical) setting. For instance, ISO 5338 on AI system lifecycle processes includes a "Stakeholder needs and requirements definition process" [ISO23c, 18]. *Stakeholder analysis* depends strongly on the individual system's context, which necessitates generalizable guidelines. In [DS22a] for instance, the authors outline a structured overview of relevant stakeholders for RAI systems. They derive a stakeholder categorization of three levels: individual (such as users, developers and domain experts), organizational (e.g. companies or institutes), and national/international (this includes law makers and regulatory agencies) [DS22a, 233]. In section 3.3.3.5, we briefly introduce stakeholder roles in alignment with the AI Act and AIRO [GD22]. Overall, this illustrated information collection is intended to function as a basis for an organized stakeholder identification process.

Human-Machine Teaming The user comprises a central and omnipresent stakeholder to be considered during design and "[...] AI system users, can be unfamiliar with AI. This can cause trust and adoption challenges" [ISO23c, 4]. Regarding the medical user for instance, findings surrounding the collaboration between physicians and intelligent systems that provide information for medical decision-making (CDSS), which the human user then translates into knowledge, are presented in [Hen22]. There, the perception of collaboration was attempted to be extracted and analyzed based on the integration of a real-time early warning system for patients with sepsis. The survey of physicians revealed that especially the teaming perspective is decisive for successful adoption of the technology. The system was perceived as a competent "second pair of eyes" [Hen22, 2], which, among other things, assists in organizational activities, such as prioritizing patient visits. Experiences from trusted sources, for instance, colleagues, and personal experiences of interacting with the system were identified as sources of trust. Knowledge of the internal functioning of the model is less crucial than providing a how-to-use approach, as physicians develop a mental model of how its influence should be evaluated during interaction with the intelligent system [Hen22, 2]. [Hen22]

Diverse Teams Multidisciplinary and diverse teams are required from the start of project conceptualization. Focusing on risk mitigation, NIST stresses the necessity of a "[...] broad set of perspectives [...]" [Tab23, 9] among contributing stakeholders, and ISO 5338 highlights the possible "[...] introduction of biases by narrow views of stakeholders" [ISO23c, 18]. Multidisciplinarity includes "[...] a diversity of experience, expertise, and backgrounds and comprise[s] demographically and disciplinarily diverse teams" [Tab23, 9]. This poses additional requirements, which need to be considered for project planning. Diverse teams require appropriate communication channels and a supportive environment to function effectively.

3.3.2.2. Domain Knowledge

The inclusion of *Domain Knowledge* for project conceptualization is crucial to ensure that the intelligent system fulfills its intended purpose in the real world, which is why it is included as a fundamental pillar for MQG4AI design, and it is further outlined in section 3.3.3.1 with respect to the individual application.

Use Case-Specificity Design decisions along the AI lifecycle are characterized by use case-specificity, which necessitates a reasonable comprehension of the respective domain and existing real world conditions. In addition to the previously introduced domain-specific impact regarding the evaluation of TAI criteria, AI risks, and concrete implementation of design decisions, this includes in-domain testing [Tab23, 15] strategies, as well as a reliable performance evaluation approach.

Evaluation in the Real World Especially in medicine, the inclusion of domain-

and use case-specific knowledge is indispensable for efficiently training and accurately evaluating the AI model. AI-based software for healthcare is primarily designed to enhance clinical treatment and patient care, and the model's intended application and objective impact its design from the start of project conceptualization. The AI Act introduces the *intended purpose* as "the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation;" in Article 3 [Fut24], which highlights a profound understanding of the intelligent system's real world context and is closely related to RM, as further outlined in the following sections on MQG4AI's contextual information blocks. Further, ISO 5338 on *AI system lifecycle processes* includes a "business or mission analysis process" [ISO23c, 17] as essential part of lifecycle management, and with respect to AI, they stress the consideration of possible risks that might complicate the achievement of certain objectives early on for a successful project [ISO23c, 18]. IS/ISO/IEC 42001 on *AI Management* equally emphasizes "[u]nderstanding the organization and its context" [ISO23b, 5], implying that relevant domain knowledge needs to be translated for different stakeholders.

Domain Knowledge Translation Defining, and preparing essential knowledge on the real world, tailored to the respective stakeholder along the AI lifecycle is not a trivial question. For instance, relevant knowledge on AI for cardiologists that are interested in actively contributing to the AI lifecycle could base their knowledge on [HW22b] and [HW22a]. They offer a comprehensive analysis of basic AI notions and use cases in cardiology. Vice versa, developers of intelligent systems that are located in cardiology need a basic understanding of the medical domain to comprehend the data they are working with, and extract other relevant knowledge that comprises the foundation for technical and conceptual design decisions, such as tuning objectives or performance evaluation metrics during model development, for instance.

3.3.2.3. Interdependency Analysis

AI lifecycle design is characterized by interdependencies and an interdependency analysis of AI lifecycle information is crucial to ensure the desired behavior in its intended real world setting. Interdependencies reach from supplementing processes such as stakeholder definition, and application analysis to more technological interdependencies between the data and the model, for instance. Aiming for a reliable impact anticipation of AI systems [Tab23, 24], *Interdependency Analysis* comprises a fundamental design principle of the MQG4AI blueprint.

Process Dependencies "The AI lifecycle consists of many interdependent activities [...]" [Tab23, 24]. ISO 5338 for instance, outlines four categories of processes

(Agreement processes, Organizational project-enabling processes, Technical management processes, and Technical processes) that consist of a multitude of subprocesses, including technical information management, and knowledge management of technical processes [ISO23c].

> In practice, [contributing stakeholders] in charge of one part of the process often do not have full visibility or control over other parts and their associated contexts. For example, early decisions in identifying purposes and objectives of an AI system can alter its behavior and capabilities, and the dynamics of deployment setting (such as end users or impacted individuals) can shape the impacts of AI system decisions. As a result, the best intentions within one dimension of the AI lifecycle can be undermined via interactions with decisions and conditions in other, later activities. This complexity and varying levels of visibility can introduce uncertainty into risk management practices. [Tab23, 24]

Aiming for risk mitigation by design, we envision flexible, organized and comprehensive information access to all contributing stakeholders along the AI lifecycle, which is based on DSR and detailed in section 3.3.7. Shared and complete information access enables the on-going extraction of complex interdependencies for MQG4AI design.

Technical Dependencies Further, identifying and monitoring technical dependencies is crucial. This is based on DNN's inherent complexity and evolutionary character, as well as our current level of AI design knowledge regarding the implementation of reliable real world behavior. Different design decisions along the intelligent system's lifecycle are related, and can result in AI pitfalls, if not anticipated, which impacts the desired behavior. To ensure accurate and trustworthy results, it is essential to comprehend the foundations of DL and address methodological pitfalls [HSA22] [TJ21] [MF22] [SQ21], as further outlined in chapter 2 for model development. Consequently, technical interdependencies need to be identified for a reliable AI lifecycle implementation.

Reliable Design Decision-Making With respect to facilitating the implementation of RAI, a generalizable classification of reliable design decision-making for (groups of) use cases would equally support quality assessment. This task is not trivial, and the identification of structural similarities from a technical viewpoint between AI use cases is considered a reasonable approach. For instance, in [SN20], the authors adapted sample-centric multi-label classification metrics from protein discovery to ECG multi-label classification [SN20, 3] based on shared model capabilities. More of possibly generalizable methods need to be identified for cumulative RAI knowledge generation, focusing on the abstraction of interdependencies. Overall, MQG4AI aims to enable comprehensive and clear design decision guidelines along the AI lifecycle in an on-going manner.

3.3.2.4. Risk Analysis & Mitigation

As introduced in section 3.2, RM plays a crucial role for high-risk domains such as medicine and is required by law. RM "[...] identifies, analyses, treats and monitors risk throughout the life cycle [...]" [ISO23c, 13]. The other three fundamental principles each highlight different nuances towards *Risk Mitigation*, which comprises the foundation of MQG4AI design. As introduced, RM is not trivial to implement for the multitude of possible use cases, and the MQG4AI information blueprint is envisioned to support RM along the AI lifecycle through formatted IM that is built on and contributes to continuous RAI KM.

3.3.3. Information Block: System

Based on the online repository of the AIRO,[17] which incorporates the AI Act and the ISO 31000 series on RM [GD22], we introduce MQG4AI's integrated information blocks that complement lifecycle design decisions. For MQG4AI blueprint design as part of this thesis, we focus on information-docking of relevant contextual information on the intelligent system and risks, highlighting interdependencies along the AI lifecycle. The proposed concept *Information Block* are envisioned to be dynamic within MQG4AI. Therefore, they are intended to be exchangeable with existing QM procedures, that possibly already collect the required information as defined in the AI Act. Thus, we aim to provide customizability of the RAI information flow and further QM information.

This section outlines AI system-information in more detail. We identified four sub-sections relevant to the AI system as a starting point: *Application*, *Compliance*, *Documentation*, and *Stakeholder* [GD22]. The next section briefly presents the lifecycle-section, which is introduced in more detail in the next chapter 4. The second identified supplementing information block on RM information is proposed in section 3.3.5, after briefly highlighting QGs along the AI lifecycle in section 3.3.4. They are detailed in chapter 4.

3.3.3.1. Application

The *Application* information block comprises relevant knowledge on the intended use case, focusing on its capabilities, purpose, and real world context, including information for impact assessment [GD22]. It is intended to provide a structured approach to understanding and organizing domain-specific and ethical information for all stakeholders. Therefore, we append an *Ethics_Specific*, and a

[17] `https://delaramglp.github.io/airo/`, which is an unofficial draft as of June 2025.

Domain Knowledge subsection aiming to provide relevant knowledge on the concrete *Application* enabling stakeholders to guide a comprehensive impact analysis and high-quality implementation of the AI lifecycle. This approach follows MQG4AI's four fundamental principles, as introduced in section 3.3.2. For input on ethical project planning, refer to section 3.2.2.3.

On a general level, the *Domain* of choice defines the intelligent system, as well as applicable regulation, which both shape the implementation. It describes the intended sector, or industry for system application, and outlines the broadest definition of the system's application, which is further specified by the following attributes. Section 3.3.2.2 introduces the relevance of *domain knowledge* in more detail. As an example how to approach healthcare, using ISO/TR 24292 [ISO21b] on *medical intelligent applications*, medical AI systems are classified by their medical specialty (capability) and medical usage (purpose). This classification informs the area of impact, with additional details on the system's output, intended workflow, and environment of use refining the broader application definition.

Further, information on the application results in a general project overview that provides the foundation for risk analysis:

- *Purpose*: The intended purpose of the AI system, such as aiding diagnosis or improving workflow efficiency refers to the intended use, as defined by the provider, and as outlined in the required documentation.

- *Capability*: Functional abilities like classification, regression, anomaly detection, generation, or object detection, enable the implementation of the system's purpose.

- *Area of Impact*: Identified based on the system's purpose and capabilities to assess its impact on the real world, including stakeholders and processes.

Additional information on the intended real world integration of the intelligent system is included:

- *Modality*: Describes the form in which the system exists in the real world, which is the foundation for user-interaction scenarios.

- *Locality of Use*: Defines where the system integrates into real world operations, and we highlight the identification of the intended workflow (see Modality).

- *AI Output*: Describes the results generated by the AI system within its intended context.

Finally, the following information on the application is closely related with AI risks that are derived from the first ALTAI criterion *Human Agency and Oversight* [Hig20]:

- *Mode of Output Controllability*: Defines how the generated output can be controlled by humans.
- *Automation Level*: Defines the degree of automation under which the AI system operates.
- *Human Involvement*: Specifies the role of human oversight and interaction with the system.

After outlining the information block on the *Application*, which contributes to the overall system information, based on the AIRO webtool, the next section briefly highlights information on measures towards compliance of the intelligent system.

3.3.3.2. Compliance

The *Compliance* section provides information on standards the systems conforms to, the regulation it complies with, as well as code of conducts the AI project follows, according to the AIRO online repository [GD22]. MQG4AI is intended to adhere with the EU AI Act by design, which includes alignment with best practices and standards. For an overview of RAI knowledge, as integrated with this version of the MQG4AI blueprint in addition to overall concept design, refer to section 3.3.1.2. Other officially accepted guidelines may be included accordingly, and part III is intended to introduce decentralized workflows. For application of MQG4AI in individual AI projects, the *Compliance* information block offers customizability to append more compliant-relevant information. Next, an overview of the required documentation is outlined.

3.3.3.3. Documentation

The *Documentation* unit lists the essential documentation requirements for the AI system, covering its data, design, testing, usage, execution environment, RM, and post-market monitoring processes, based on the AIRO web tool. It incorporates relevant documentation, as required by the AI Act. [GD22] An approach how to organize the required documentation through AI cards can be found in [Gol24], which is equally oriented towards the AI Act. The MQG4AI lifecycle blueprint is intended to provide the required information, which can be extracted, transformed, and updated into document-format, next.

- *Data* documentation includes a description of data characteristics such as its distribution, or a specification of input data using datasheets. Further reasonable information extraction regarding the data could be based on the *IEEE Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence* [IEE22], for instance.

- *Overall System* information that needs to be documented comprises documentation of the system's blueprint, including the internal layout of the product the system is part of. Further, system design specification and system architecture documents need to be included. This can be based on MQG4AI. Finally, an EU declaration of conformity is required.

- *Execution Environment* documents include information on AI hardware specifications, as well as hardware the system runs on and an overview of applied components/tools. Though not part of our contribution, the MQG4AI blueprint could be extended with relevant information.

- *Post-Market Monitoring* documentation comprises a description of the required post-market monitoring system, which can be based on the MQG4AI information structure.

- *Risk Management System* analogously to post-market monitoring, the RMS information can be based on the MQG4AI blueprint.

- *Testing* requires test logs and test reports documenting system evaluations, which can equally be based on different MQG4AI template versions during individual AI projects for concept and output traceability of lifecycle versions, as outlined in section 3.3.7.3 in more detail.

- *Use* documentation includes installation instructions, as well as instructions for use. The latter are required to detail information such as acceptable accuracy levels and the interpretation of relevant metrics, information which can be derived from MQG4AI.

3.3.3.4. Ethics

The *Ethics_General*[18] module is intended to provide a fundamental overview of ethics in the context of RAI, designed to equip all stakeholders with essential knowledge and foster a RAI mindset. Recognizing the human mindset as a critical component of the AI lifecycle, this unit addresses related risks through offering foundational guidance for ethics training. For ethical questions tailored

[18] https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/1_System/Ethics_General/Ethics_General.md

to specific applications, refer to *Ethics_Specific* in section 3.3.3.1. This interplay between knowledge on general ethical principles and their manifestation in the context of individual use cases is equally proposed in [GC20], where the author introduces an ethical framework to guide the intelligent system's actions in its intended real world setting. Basic information on *General Ethics* equally provides the foundation for the development of ethics training content. We outline exemplary starting points for ethical considerations in the context of RAI in [EM25b], where we propose a holistic RAI mindset. A key guiding question to design this system-related section is: *What ethical information on risks and trustworthiness should be distributed to all stakeholders?*

The human influence on AI underscores the importance of cultivating a RAI mindset, as introduced in section 3.2.2.3. This aligns with the ethical framework of the EU AI Act, rooted in the Assessment List for Trustworthy AI (ALTAI) framework published in 2019 by the European Commission's High-Level Expert Group on Artificial Intelligence [Hig20]. ALTAI categorizes ethical AI qualities into seven key areas that are outlined further in section 3.2: Human Agency and Oversight, Technical Robustness and Safety, Privacy and Data Governance, Transparency, Diversity, Non-discrimination and Fairness, Societal and Environmental Well-being, and Accountability. These categories form the basis of our proposed holistic approach, aiming to shape a RAI mindset among stakeholders. By examining risks through these ethical lenses, stakeholders can enhance their ability to identify and evaluate risks effectively. Furthermore, this approach emphasizes the indispensable role of research and education in promoting awareness of RAI.

Finally, the four fundamental *MQG4AI Design Principles* towards risk mitigation, as introduced in section 3.3.2 complement the creation of a RAI mindset aiming towards applicability, contributing to embedded ethics [MS20]. These principles serve as actionable "answers" to support stakeholders in conducting thorough risk analyses, including both identification and evaluation. Stakeholders are encouraged to reflect on these principles and consider their interpretation and application in their specific contexts: *How are these principles interpreted by you?*

3.3.3.5. Stakeholder

The *Stakeholder* element within the *Application* information block is intended to provide an overview of stakeholder roles for project execution and management of requirements to their roles, including required capabilities and available personnel. The, in this section introduced roles are based on the AIRO webtool, aligning with the AI Act's terminology [GD22]. Section 3.3.2.1 emphasizes the importance of *Stakeholder Inclusion* in more detail, including considerations on how to approach the definition of relevant roles.

Generally, ISO 22989 on *AI concepts and terminology* defines *stakeholder* as "any individual, group, or organization that can affect, be affected by or perceive itself to be affected by a decision or activity" [ISO22a, 12], which is included in the AIRO. *Contributing Stakeholder*, as referred to in this document comprise those who "affect" the AI system's "decision or activity", as previously introduced in section 3.2.2.3.

In addition, we propose to globally structure stakeholders into *active*, *consulting*, and *passive* roles for directed IM, since the stakeholder's involvement with the project impacts how information should be tailored and presented to them. *Contributing Stakeholder* can comprise active (e.g. developer), consulting (e.g. domain expert), and passive (e.g. user) roles depending on how the intelligent system is implemented. Overall, stakeholders impact the lifecycle implementation from various perspectives, as well as pose or mitigate AI risks. For instance, this is articulated by the fifth ALTAI requirement *Diversity, Non-Discrimination, and Fairness - Stakeholder Participation* [Hig20] in section 3.3.5.5. The following provides a generic overview of stakeholder roles, as summarized in the AIRO online repository [GD22], which references Article 3 in the AI Act [Fut24], including examples.

Article 3 of the AI Act summarizes a combination of roles as *AI Operator*, which is incorporated in the AIRO. The term "'operator' means a provider, product manufacturer, deployer, authorised representative, importer or distributor;" as defined in Article 3 of the AI Act [Fut24]. In the AIRO, the sub-roles *AI provider* and *AI deployer* are highlighted explicitly, and they append *AI developer* based on ISO 22989. This results in the following list of stakeholders comprising *active* and *consulting* roles:

- *AI Provider* "[...] means a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge;" [Fut24]. Among others, these roles include AI developer (develop the model) and domain experts (real world information), data scientists (preprocess and analyze data), mathematicians (the foundations of DL), as well as DevOps, and MLOps (infrastructure for continuous deployment and maintenance) that work for a particular company. Further, relevant roles comprise company executives (decide on AI investments), or quality assurance manager (oversee RM and system performance). These roles apply to other stakeholder groups, such as deployer or manufacturer, as well, depending on their responsibilities. Finally, aiming to foster the integration of ethics, possibly a TAI and/or RAI expert role could be a reasonable addition.

- *AI Developer* "An organisation or entity that is concerned with the development of AI services and products" [ISO22a, 33]. This definition is based on ISO 22989, and belongs to *AI provider*, as defined in the AI Act. They exemplify the following four sub-roles summarizing required proficiency: *model designer* (i.e. the entity responsible for receiving data and a problem specification to create an AI model); *model implementer* (i.e. the entity that takes an AI model and determines the specifics of its computational execution, including the choice of implementation and compute resources); *computation verifier* (i.e. the entity that ensures the computation is executed as intended); *model verifier* (i.e. the entitiy that monitors the model's performance in alignment with its intended purpose). [ISO22a, 33]

- *AI Product Manufacturer* is not explicitly defined in Article 3 of the AI Act, but in ISO 13485 on *Medical devices — Quality management systems — Requirements for regulatory purposes*: "[...] natural or legal person with responsibility for design and/or manufacture of a medical device with the intention of making the medical device available for use, under his[/her] name; whether or not such a medical device is designed and/or manufactured by that person himself or on his[/her] behalf by another person(s)" [ISO16, 3]. Consequently, the manufacturer plays a role, when the AI system is embedded within a product, such as a medical device. Regarding the AI system, the manufacturer for example mirrors an AI provider, if the system is developed in-house, or an AI distributor, when incorporating a model that was developed by another entity.

- *AI Deployer* "[...] means a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity; Related: Recital 13" [Fut24]. For instance, deployer of AI systems describe those, who enable interaction with the intelligent system for others. This could be a hospital that equips medical personnel with an AI tool for medical imaging, or a school that applies GenAI tools.

- *Authorized Representative* "[...] means a natural or legal person located or established in the Union who has received and accepted a written mandate from a provider of an AI system or a general-purpose AI model to, respectively, perform and carry out on its behalf the obligations and procedures established by this Regulation;" [Fut24]. This role comprises company executives that are responsible for the quality of the intelligent system, e.g. the chief technology officer (CTO), or chief executive officer (CEO), depending on the company's internal structure.

- *AI Importer* "[...] means a natural or legal person located or established in the Union that places on the market an AI system that bears the name or trademark of a natural or legal person established in a third country;" [Fut24]. AI importer work with intelligent systems that are developed outside the

EU. For instance, this comprises AI-powered chat-bots that are based on OpenAI's GPT-series, which are developed in the US.

- *AI Distributor* "[...] means a natural or legal person in the supply chain, other than the provider or the importer, that makes an AI system available on the Union market;" [Fut24]. AI distributors sell AI systems that are developed by other entities, such as a medical device that incorporates an intelligent system.

Passive roles with respect to system development comprise stakeholders that are impacted by the AI output in the real world. The following sub-groups are identified:

- *AI User* comprises an "[i]ndividual or group that interacts with a system", as globally defined in the AIRO webtool [GD22]. ISO 22989 specifies AI users as "an organization or entity that uses AI products or services" [ISO22a, 34] as a subgroup of "AI customer" that include the indirect provision of AI to AI users [ISO22a, 34]. This definition aligns more with *AI deployer*, as previously introduced and highlights differences in perspective between ISO and the AI Act. Possibly, since deployers of high-risk AI have to fulfill obligations, as outlined in Article 26 of the AI Act [Fut24]. Examples include those who interact actively and knowingly with the intelligent system, such as employees, medical personnel using intelligent enhancements during diagnosis or for clinical decision support, as well as individuals consulting LLMs for various purposes.

- *AI Subject* refers to "[a]n entity that is subject to or impacted by the use of AI", according to the AIRO webtool [GD22]. This notion results in a shift from the active user role to a more passive perception. *AI subjects* comprise by the intelligent system affected people, as well as those who are subjected to interact with the intelligent system. Continuing the examples mentioned above, this comprises patients that are diagnosed or treated using AI-supported systems, or students who use AI tools based on their school's policy. ISO 22989 further distinguishes AI subjects into *Data subjects*, and *Other subjects*, highlighting users that train the AI model, and those who interact with AI indirectly, through e.g. recommendation systems [ISO22a, 34]. The AI Act contributes an additional perspective focusing on testing of intelligent systems, and defines "'subject', for the purpose of real world testing, means a natural person who participates in testing in real world conditions;" in Article 3 [Fut24]. This perspective contributes a more *active* or *consulting* stakeholder role.

Finally, addressing the interface with regulation, notified bodies comprise important additional *active* and *consulting* roles. *Notified Body* is defined as "[...] a conformity assessment body notified in accordance with this Regulation and other relevant Union harmonisation legislation;" [Fut24] according to Article 3 in the AI Act. *Conformity assessment* "[...] means the process of demonstrating whether the requirements set out in Chapter III, Section 2 relating to a high-risk AI system have been fulfilled;" [Fut24]. Therefore, a *notifying authority*, which "[...] means the national authority responsible for setting up and carrying out the necessary procedures for the assessment, designation and notification of conformity assessment bodies and for their monitoring;" [Fut24] has appointed a general *conformity assessment body* "[...] that performs third-party conformity assessment activities, including testing, certification and inspection;" [Fut24] to become notified. Examples for German notified bodies assessing the medical domain include *MDC Medical Device Certification GmbH*, *DEKRA Certification GmbH*, or *TÜV SÜD Product Service GmbH*.[19]

After establishing relevant information on the AI system block, highlighting stakeholder roles, documentation obligations, as well as application-specific information, the next section briefly introduces the heart of MQG4AI: the generic and customizable AI lifecycle information blueprint.

3.3.4. Information Block: *Quality Gates* along the AI Lifecycle

Quality Gates (QG) along the AI lifecycle comprise the center of MQG4AI information linking and structuring for comprehensive lifecycle planning and RAI KM. The proposed generic lifecycle concept, as well as the two QG types *collection-QG*, and *leaf-QG* are elaborated in more detail in chapter 4.

Overall, the concept QG is derived from software quality management and product development practices. It broadly describes "[...] an objective quality assurance gate, that is, a verification procedure, performed either by independent reviewers or by automated scripts" [Fil06, 34]. Their fundamental role is to consolidate key criteria related to specific outcomes produced at various stages of the software development or product lifecycle [Flo08, 245]. [GM09] [HS09] [Fil06] [Flo08].

In this thesis, we adapt this high-level definition to the AI context by focusing on AI lifecycle design decisions, which are driven by design knowledge and monitored through information. We propose organizing and processing relevant choices using QGs, instantiated through a specific QG design decision-making framework and a QG identification process, which both mirror key criteria for quality assurance. This approach operates on the premise that transparent AI lifecycle design choices lead to achieving the desired quality.

[19] For more information consult the notified bodies repository by the European Commission: `https://webgate.ec.europa.eu/single-market-compliance-space/notified-bodies`

In general, QGs design the lifecycle from generalizable to use case-specific implementation guidelines, and their generic and customizable information processing format is intended to address the dynamic nature of AI through IM within the MQG4AI lifecycle blueprint, including evolving versions, and contextual information docking. The overall aim is to create a dynamic, generalizable and application-oriented lifecycle skeleton through decentralized QG identification workflows for continuous MQG4AI design. We simultaneously aim to pave the way for RAI implementations of individual AI use cases that benefit from existing RAI design knowledge. The lifecycle planning blueprint is envisioned as an IM structure that functions a tool for contributing stakeholders to continuously interact with the AI lifecycle concept as a team.

In addition to highlighting technical interdependencies, as outlined in section 3.3.2.3, this includes the identification of connections between design decisions for implementation with supplementary contextual information. For instance, the intelligent system and AI risks impact and are impacted by lifecycle design choices, and we aim for a reliable integration of the intelligent system with the real world through continuous lifecycle planning that results in robust and adaptable system behavior. After introducing the AI system and AI lifecycle information blocks of the MQG4AI blueprint, the next section illustrates a possible structure for the RM information block.

3.3.5. Information Block: Risk Management

RM, which is introduced in section 3.2, comprises a central contextual information block of the MQG4AI blueprint. Overall, MQG4AI design envisages compliance with the EU AI Act, focusing on continuous IM of RAI knowledge along the AI lifecycle. Generally, intelligent systems are required to respect and realize fundamental rights, safety, and health [Fut24] when they are integrated with their intended environment in the real world, which is closely tied to their intended purpose and functionality. This is detailed in section 3.1 on RAI. Especially, high-risk systems, if not implemented accordingly, pose risks to this vision.

As established in section 3.2.2.2, this section introduces the "Ethics guidelines for trustworthy AI" [Eur19] in more detail to provide an overview for AI risk classification through TAI criteria. The related guiding questions, as outlined in the ALTAI [Hig20] are included with the MQG4AI lifecycle blueprint, and we highlight for a selection of criteria, how they are addressed within MQG4AI to contribute to risk mitigation, which comprises the essence of MQG4AI design. It should be noted that future European AI standardization, whose key areas align with TAI criteria, is intended to prioritize addressing identified risks [SG24b, 4]. Therefore, the application of European standards, which are currently in their infancy surrounding AI, as introduced in section 3.1.3.2 functions as a foundation to identify relevant information for the RM process that can be linked with the AI lifecycle within MQG4AI. The next section highlights the estimated information flow between the three introduced MQG4AI information blocks (*System*, *Lifecycle*, *RM*), and section 3.3.7 outlines the underlying method (DSR [vJ20]) to enable decentralized and continuous RAI KM, which includes AI risk classification with respect to the implementation of risk controls along the AI lifecycle.

Perspectives on RM Overall, risks are considered from multiple perspectives, and run through the entire blueprint, highlighting the, in section 3.3.2 identified four fundamental design principles towards risk mitigation. Section 3.2 introduces basic RM processes, which contribute information to the RM module. We briefly introduce the OECD interoperability framework [Org23], and the NIST AI RMF [Tab23], and discuss their alignment with the European perspective on RM and TAI, as outlined in Section 3.2.2. In addition to required documentation and application assessment, as introduced in sections 3.3.3.1, and 3.3.3.3, risk mitigating measures comprise shedding light on the design decisions that build the AI lifecycle, thereby enhancing transparency.

Use Case-Specificity Challenges related to AI risks and their control arise from their use case-specific nature, their evolving characteristics, and the limited practical experience we currently have with intelligent system behavior in real world environments. The *AI Risk Repository*[20], for instance provides over 700 identified risks, which comprises a solid foundation for risk identification. Focusing on AI security and threats, the *Mitre Atlas*[21] offers a "living knowledge base" for continuous identification. Both repositories serve as a starting points for addressing known AI risks across the AI lifecycle. Another approach to addressing the use case-specific nature of AI risks is presented in [LS25], where the authors introduce the RAI "[...] Question Bank [...] designed to support diverse AI initiatives" [LS25, 1] in identifying AI risks.

Linking along the AI Lifecycle Overall, we envision closing the gap from abstract to use case-specific RAI design decision-making for comprehensive AI lifecycle planning based on MQG4AI. This includes a structured and generic approach to AI risk identification within the MQG4AI blueprint to enable a comprehensive and continuous information-docking towards a risk mitigated AI lifecycle implementation during individual AI projects, as well as the contribution of design knowledge on how to implement risk control measures along the AI lifecycle. Information on AI risks is linked with the AI lifecycle implementation through QGs, which is demonstrated in more detail in chapter 4.

The following sections outline TAI [Hig20] for risk identification, so that RM can be directly linked with the AI lifecycle to monitor implemented risk controls, which is exemplified in part III through MQG4AI interaction scenarios.

[20] https://airisk.mit.edu/

[21] https://atlas.mitre.org/

3.3.5.1. Human Agency & Oversight

Human Agency & Oversight comprise one category to identify, collect, and monitor AI risks related to the intelligent system's interaction with humans, and the human involvement, which is subdivided into two sub-categories according to ALTAI [Hig20]. "AI systems should support human agency and human decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both: act as enablers for a democratic, flourishing and equitable society by supporting the user's agency; and uphold fundamental rights, which should be underpinned by human oversight" [Hig20, 7].

- ***Human Agency & Autonomy*** focuses on "[...] the effect AI systems can have on human behaviour [...]" [Hig20, 7], which needs to be anticipated from the start of project conceptualization for RAI implementations. For instance, this includes the consideration of *automation bias*, or the tendency of human users to rather trust the system's output than to critically second guess it [ISO21a], which impacts the design of the system's user interaction interface, as well as an on-boarding strategy that includes user training. Other negative consequences can directly target end users, e.g. by creating an addictive system dependency [Hig20, 8]. Also, this criteria requires a clear communication of the AI output to avoid confusion [Hig20, 7].

- ***Human Oversight*** functions as a risk mitigating measure "[...] through governance mechanisms [...]" [Hig20, 8], to ensure that "[...] an AI system does not undermine human autonomy or causes other adverse effects" [Eur19, 16]. Generally four different modes are identified, and "[...] the less oversight a human can exercise over an AI system, the more extensive testing and stricter governance is required" [Eur19, 16]. The fitting approach for the individual use case needs to be identified:

  - self-learning, autonomous systems act without human intervention.

  - *Human-in-the-Loop* means that a human being can intervene in every AI output for individual data samples.

  - *Human-on-the-Loop* covers human intervention in the system's design and runtime oversight.

  - *Human-in-Command* is able to oversee an AI system's activity, including its broader impacts, and decide when, how, or if to use it, including the ability to override a decision made by an AI system. [Hig20, 8]

Chapter 8 introduces a *human-in-the-loop* approach, aiming to illustrate MQG4AI application scenarios, as outlined in section 3.3.7.3, analyzing related risks.

3.3.5.2. Technical Robustness & Safety

"*Technical robustness* requires that AI systems are developed with a preventative approach to risks and that they behave reliably and as intended while minimizing unintentional and unexpected harm as well as preventing it where possible. This should also apply in the event of potential changes in their operating environment or the presence of other agents (human or artificial) that may interact with the AI system in an adversarial manner." [Hig20, 9] To achieve RAI, systems must be dependable (justifiably trusted) and resilient (robust to changes), which is based on a secure and reliable AI lifecycle.

- ***Resilience to Attack and Security*** covers risks related to the protection against vulnerabilities to AI-specific cyber-attacks such as *data poisoning*, when training data is manipulated, *model evasion*, when the model's output is influenced, or *model inversion*, when model parameters are changed. [Hig20, 9] The underlying infrastructure, including hardware and software components such as specific libraries, may be vulnerable to attacks, and potential weaknesses must be identified. [Eur19, 16]

- ***General Safety*** comprises requirements that refer to the RM system and RM processes in general. This includes the definition and quantification of risks, identification of risk sources in alignment with possible misuse scenarios, as well as how AI output is related with stable or unstable behavior [Hig20, 10].

MQG4AI blueprint design is intended to support this process through linking AI risks with the lifecycle implementation, as deepened in chapter 4. This module can be interpreted to function as an overview to assess the quality of the implemented RMS through the identification of related risks.

- ***Accuracy*** AI output correctness is crucial to guarantee the reliability of actions that are based on the intelligent system's predictions. This requires "[a]n explicit and well-formed development and evaluation process [...]" [Eur19, 17], as well as measures to continuously monitor and document the system's performance [Hig20, 10].

MQG4AI is intended to support this process through information extraction and transformation along the AI lifecycle through QGs, as outlined in chapter 4. The related risk *unreliable performance evaluation metrics*, and how to organize implemented risk controls within MQG4AI is demonstrated in chapter 7.

- ***Reliability, Fall-back plans and Reproducibility*** Results need to be reliable and reproducible, which describes an intelligent system "[...] that works properly with a range of inputs and in a range of situations" [Eur19, 17].

This comprises verification and validation procedures, fallback plans if errors occur, as well as monitoring if the "AI system is meeting the intended goals" [Hig20, 11].

QGs are intended to support this process, through enabling flexible, comprehensive and continuous lifecycle management, as introduced in more detail in the next chapter 4.

3.3.5.3. Privacy & Data Governance

Data plays a crucial role when designing and implementing AI systems, and "[t]he quality of the data sets used is paramount to the performance of AI systems" [Eur19, 17]. In addition, especially, in sensitive settings such as medicine, respecting privacy is essential. For instance, leaking information on the medical condition of a patient to the employer might result in unfair treatment. "Closely linked to the principle of prevention of harm is privacy, a fundamental right particularly affected by AI systems. Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy" [Hig20, 12].

- ***Privacy*** Guaranteeing privacy is closely related to trust in the system and includes risks related to the data collection process, and how it is used and accessed once collected [Eur19, 17].
- ***Data Governance*** covers data quality-related questions, as well as how the data is handled throughout the complete lifecycle of the intelligent system, aiming for continuous data protection [Eur19, 17].

3.3.5.4. Transparency

Further aiming to enhance trust in the intelligent system are requirements related to *Transparency* through the mitigation of related risks. "This requirement is closely linked with the *principle of explicability* and encompasses transparency of elements relevant to an AI system: the data, the system and the business models" [Eur19, 18].

- ***Traceability*** is intended to address related risks that result from inadequate documentation and monitoring of lifecycle design decisions, which includes mapping the model with the data is was trained on, i.e. the inclusion of data versioning techniques, and "[...] measures to continuously assess the quality of the output(s) of the AI system" [Hig20, 14].

Documenting the evolving lifecycle concept through QGs aims to support related risk mitigation, as outlined in more detail in the next chapter 4.

- ***Explainability*** DNNs are complex systems that are not comprehensible to human beings by design. "In those circumstances, other explainability measures (e.g. traceability, auditability and transparent communication on the AI system's capabilities) may be required, provided that the AI system as a whole respects fundamental rights. The degree to which explainability is needed depends on the context and the severity of the consequences of erroneous or otherwise inaccurate output to human life" [Hig20, 15]. Refer to section 2.2.2 for more information on XAI methods that support shedding light on the model's inner workings from multiple angles.

In chapter 6, we propose the design of a reliable, generic and risk mitigated explanation stage during model development within MQG4AI. We illustrate an generalizable design workflow through the transformation of related best practices to risks, and their linking with the lifecycle through QGs.

- ***Communication*** collects risks that arise from an insufficient communication of the AI system and consequences of its application to users, who should always be aware of the fact that they are interacting with an intelligent system [Eur19, 18].

3.3.5.5. Diversity, Non-Discrimination & Fairness

Inclusion and diversity are crucial components along the RAI lifecycle [Eur19, 18]. "AI systems, during both training and operation, can inadvertently reflect historical biases, incomplete data, or poor governance, leading to unintended prejudice and discrimination. Discriminatory biases should be identified and addressed during data collection. AI systems must be user-centric, ensuring accessibility for all individuals, regardless of age, gender, abilities, or characteristics, with particular attention to the needs of persons with disabilities." [Hig20, 16]

- ***Avoidance of Unfair Bias*** Related risks are closely tied with data quality, while this module highlights criteria centered around fairness. The definition of fairness is not a trivial question and depends heavily on the intended context and application of the intended system in the real world, which shapes related risks for individual intelligent systems. Consequently, contributing stakeholders need to be aware of possible biases that result in unwanted behavior, which necessitates the inclusion of ethics during project planning, as introduced in section 3.1.2.

- ***Accessibility and Universal Design*** With respect to fairness for user interaction with intelligent systems, they need to be designed in an inclusive manner that respects "[...] the variety of preferences and abilities in society" [Hig20, 17].

- ***Stakeholder Participation*** Finally, as outlined in more detail in section 3.3.2.1 on stakeholder inclusion, "[...] the widest range of possible stakeholders in the AI system's design and development" [Hig20, 18] need to be considered for a trustworthy application.

We exemplify the related risk *Lack of Domain Experts and Collaboration Mechanisms* in the MQG4AI blueprint, and outline how it corresponds to reliable performance evaluation metrics in chapter 7.

3.3.5.6. Societal & Environmental Well-being

AI is a disruptive technology that can change the world as we know it. In recent times, intelligent systems are more and more entering the real world from various perspectives and domains, which requires a holistic approach for impact assessment and monitoring of anticipated outcomes towards RAI implementations. "Social AI systems, pervasive in education, work, care, and entertainment, may reshape perceptions of social agency or negatively affect relationships and attachment. While AI can enhance social skills, it also risks their deterioration, potentially harming physical and mental well-being. The impacts of AI must be carefully monitored and addressed. Sustainability and ecological responsibility in AI should be promoted, alongside research into AI solutions tackling global challenges, such as the Sustainable Development Goals[22]. Ultimately, AI should serve the well-being of all people, including future generations, fostering democratic processes and respecting diverse values and life choices. AI systems must not compromise democracy, human deliberation, voting systems, or pose systemic risks to society" [Hig20, 19].

- ***Environmental Well-being*** Resources on our planet are limited, and especially, with respect to future generations and nature, a sustainable consumption is crucial. Challengingly, AI systems are characterized by high energy consumption, reaching from training to hardware production, and

[22] `https://sdgs.un.org/goals`

"[m]easures securing the environmental friendliness of AI systems' entire supply chain should be encouraged" [Eur19, 19].

MQG4AI's information blocks combined with transformation layers within individual QGs, as introduced in the next chapter 4, are designed to enable customizable extraction of additional information. While we illustrate this mechanism in the context of RM in general, it can equally be applied to other areas of relevant QM information, or specify certain topics, such as assessing resource consumption for measuring sustainability across the AI lifecycle.

- ***Impact on Work and Skills*** It is not yet clear, how interacting with AI systems will impact human behavior, and effects of intelligent systems "[...] must therefore be carefully monitored and considered" [Eur19, 19].

- ***Impact on Society at large or Democracy*** Society is continuously evolving, and technological impact influences the direction of change, and "[...] impact should also be assessed from a societal perspective" [Eur19, 19] in relation to intended use and misuse scenarios.

3.3.5.7. Accountability

The implementation of *Accountability* complements to ensuring the other criteria [Eur19, 19]. As a consequence, RM functions as a means to implement TAI, and contributes towards implementing RAI, as introduced in section 3.1, through the provision of guidance for lifecycle design decision-making. MQG4AI is intended to offer a comprehensive concept that supports AI QM, which includes RM and enables the implementation of RAI systems, that promote "[...] auditability and accountability during [their] design, development and use, according to specifications and the applicable regulation of the domain of practice in which the AI system is to be used" [DRN23, 18]. Or, in other words, "[t]he principle of accountability requires mechanisms to ensure responsibility for the development, deployment, and use of AI systems. It is closely tied to transparent risk management, enabling risks to be identified, mitigated, explained, and audited by third parties. In cases of unjust or adverse impacts, accessible accountability mechanisms should provide options for redress" [Hig20, 21].

- ***Auditability*** The implemented AI system should be designed in a way that enables quality assessment of "[...] algorithms, data and design processes" [Eur19, 19], which is closely tied to transparent IM. High-risk intelligent systems "[...] should be able to be independently audited" [Hig20, 21].

MQG4AI design is intended to support *auditability*, or the conformity assessment process through continuous interaction with a comprehensive RAI lifecycle blueprint, which is introduced in more detail in the next chapter 4.

- ***Risk Management*** In addition to the *General Safety requirement* in section 3.3.5.2 that focuses more on the quality of RM processes such as risk identification, this unit concentrates on surrounding mechanisms, such as the establishment of an AI ethics review board, third-party auditing processes, or processes for continuous monitoring. Failing to establish such processes in a qualitative manner creates risks that can compromise the effective integration of intelligent systems into real world applications. In this context, highlighting handling of trade-offs between TAI criteria is important, which is characterized by high use case-specificity [Eur19, 20].

MQG4AI design is intended to support on-going RM through continuous and comprehensive lifecycle planning based on different versions that render the evolution of design decisions traceable. This is realized through different MQG4AI interaction scenarios, as introduced in more detail in section 3.3.7, and chapter 4. After introducing the proposed information structure for the RM module within MQG4AI, the next section summarizes the identified information flow between the three exemplified blueprint information blocks.

3.3.6. Inter-sectional Information Flow

This section illustrates the extracted information flow between the three previously outlined MQG4AI information blocks, before we introduce MQG4AI interaction scenarios based on DSR [vJ20] to enable continue designing the blueprint in a decentralized manner, in the next section. MQG4AI comprises three primary modules: *System*, *Lifecycle*, and *RM* information, with an interconnected, bidirectional information flow among them. Each unit provides relevant insights to the others, forming a cohesive IM structure that focuses on the extraction of interdependencies of relevant information for RAI lifecycle conceptualization. These modules were identified to prioritize RM, ensuring a reliable and robust lifecycle implementation of AI systems. The information flow between these information blocks is depicted in Figure 3.15, and exemplified as follows.

System Information

- **Shapes the Lifecycle Implementation** The application's intended purpose and capabilities, along with domain knowledge, are crucial for shaping individual lifecycle implementations and ensuring reliable design decisions. For instance, when designing a CDSS to assist doctors with diagnosis based on medical imaging, knowledge on medical imaging standards such as DICOM, or disease-specific biomarkers shape design decisions.

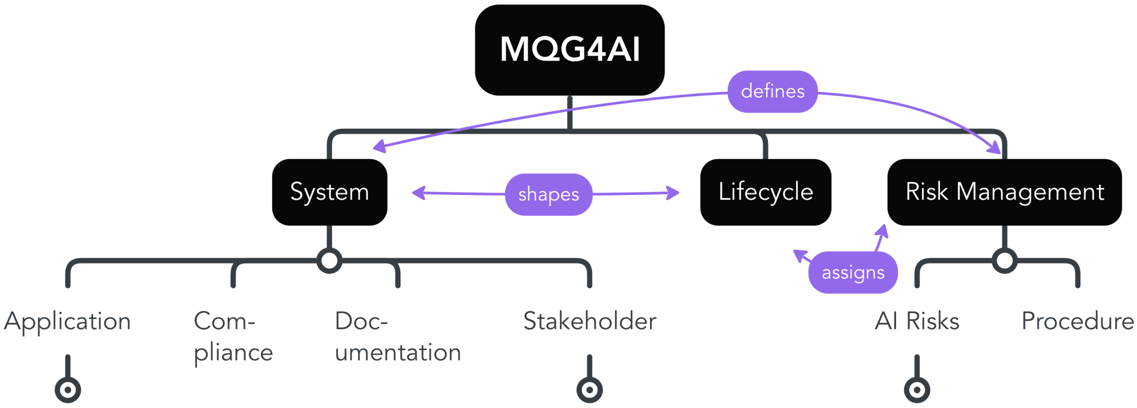


Figure 3.15.: Information flow between MQG4AI information blocks.

- **Defines RM Information** The application's area of impact helps identify potential risk sources for analysis. Also, identified stakeholder roles inform risk analysis, when anticipating system misuse scenarios. For instance, remote patient monitoring systems need to communicate critical events such as heart abnormalities without delay, and with a high precision, otherwise, medical personnel might get accustomed to the system's signals, which could result in system misuse through silencing or ignorance.

Information on the Lifecycle Implementation

- **Shapes System Information** For instance, the post-market monitoring system needs to be documented, which includes the extraction of events that trigger system updates when data drift occurs and the system's performance is impacted. This monitoring and update process must be documented, outlining the specific conditions that prompt system modifications, such as the correct interpretation of threshold deviations. Also, information on the system's accuracy metrics needs to be communicated clearly to the user in the instructions of use. Users are also informed about potential limitations, such as decreased accuracy for individuals with specific conditions that were underrepresented in the training data.

- **Assigns Information to RM** A compilation of or individual design decisions may function as risk controls or risk sources. For instance, a small data set, which is common in medicine due to high labeling costs and complex privacy rights, may increase the risk of overfitting to the available data, which negatively impacts the desirable generalizability of the model. This necessitates additional design decisions for risk mitigation, such as transfer learning and a human-in-the-loop approach.

Risk Management

- **Defines System Information** The RMS needs to be documented throughout the lifecycle and its evolutions. Identifying and monitoring relevant information is part of the MQG4AI blueprint IM structure, which links risks and risk controls with the lifecycle implementation in a structured way for a transparent design decision-making process. Further, the implementation of risk controls, such as the consultation of an AI ethics review board [Hig20, 22] shapes stakeholder roles. Other risk controls include ethics training for stakeholders on TAI tailored to the specific application, or encouraging continuous learning for all contributing stakeholders to enable continuous RAI knowledge updates.

- **Assigns Information to the Lifecycle Implementation**. The implementation of risk controls needs to be continuously monitored, which necessitates observing generated results, and (re-)evaluating residual risks. Continuous risk control measures can include the establishment of feedback loops for users to report unexpected system behavior, or regular testing protocols that need to be implemented.

Overall, the MQG4AI blueprint aims to provide a unified approach for comprehensive and continuous lifecycle planning towards implementing RAI. It is based on the premise that every lifecycle design decision contributes information to achieving the desired quality of AI systems. Transparent information-extraction and -linking forms the core of MQG4AI. This on-going information collection is envisioned to function as foundation for the required information to document, as outlined in more detail in [Gol24] summarized as "Model Cards". After introducing the MQG4AI information blueprint in the previous sections, the next section presents the envisioned interaction scenarios for application and knowledge contributions towards RAI by design, guiding the lifecycle from generalizable to use case-specific implementation guidelines, under consideration of on-going evolutions in a dynamic real world.

3.3.7. Towards Decentralized & On-Going MQG4AI Interaction Scenarios for Evolving RAI Knowledge

Rounding up the conceptual introduction of MQG4AI, this section focuses on MQG4AI's design method *Design Science Research* (DSR) [vJ20], which comprises the foundation for developing the continuously evolving RAI lifecycle planning and monitoring blueprint. First, DSR is introduced, next, we outline two perspectives that are envisioned to enable future contributions in a continuous manner through decentralized MQG4AI interaction scenarios, before demonstrating how the current version of the MQG4AI lifecycle blueprint is designed based on a retrospective analysis of the use cases presented in part III. The extraction of a generalizable design workflow contributes to our vision RAI, and the next section

3.4 summarizes long-term objectives and limitations of our proposed approach, before diving deeper into MQG4AI's setup and components in chapter 4.

3.3.7.1. Method *Design Science Research*

DSR [vJ20] is selected as a suitable method for RAI knowledge generation and application, since it corresponds to the evolutionary character of RAI design knowledge and the technology's use case-specificity through the provision of mechanisms for dynamic knowledge interactions with the real world. For a visual explanation, see Figure 1.1 in section 1.4.

> Design Science Research (DSR) is a problem-solving paradigm that seeks to enhance human knowledge via the creation of innovative artifacts. Simply stated, DSR seeks to enhance technology and science knowledge bases via the creation of innovative artifacts that solve problems and improve the environment in which they are instantiated. The results of DSR include both the newly designed artifacts and design knowledge (DK) that provides a fuller understanding via design theories of why the artifacts enhance (or, disrupt) the relevant application contexts. [vJ20, 1]

Dynamic, Scattered RAI *Knowledge Base* & Diverse Application *Environments* The RAI *Knowledge Base* is constituted by publications, standards and guidelines on how to implement AI systems in a responsible manner, addressing the interface of implementation and regulation. Currently, a lot of RAI knowledge exists, scattered around the world/internet, and one criterion towards implementing RAI is to organize that knowledge in a way so that it supports general applicability for AI regulators and providers alike. The multitude of contributions and implementation approaches are tested on and derived from application to concrete use cases that are characterized by their respective *Environment*, which in turn enables empirically derived novel design knowledge. AI application scenarios are as versatile as human creativity, and different settings need different design choices that depend on complex interdependencies, as introduced in section 3.3.2.3.

Requirements for *Design* Consequently, an approach to design RAI knowledge for concrete environments needs to be based on a generic access point that is applicable to as many use cases within different *Environments* as possible. In addition, the identification of structural similarities to organize the *Design Knowledge Base* so it can be linked along the generic access point for application is desirable. Also, it is likely that novel RAI design knowledge generation will never stop, as the technology continues to advance in a dynamic real world, which necessitates continuity, and favors a decentralized approach to manage the *Design Knowledge Base*.

MQG4AI as Enabler for Continuous *Design* The MQG4AI lifecycle blueprint is envisioned as a tool to realize the *Design* step that closes the gap to apply knowledge in the real world, while simultaneously enabling continuous KM. With respect to AI, we are convinced that a comprehensive lifecycle planning information structure, that, similar to the *digital twin* concept from the industry, monitors the implemented AI system from a conceptual viewpoint, is a reasonable approach to achieve the desired quality through continuous IM and KM. Therefore, the generic and customizable lifecycle planning blueprint is designed to incorporate RAI knowledge within the proposed IM structure. This is realized through the setup of the identified supplementary contextual information blocks, and MQG4AI's four fundamental design principles, as introduced in the previous sections. In addition, QGs that reflect the AI lifecycle from generic to use case-specific design decisions at the core of MQG4AI, provide a unifying access point for the multitude of possible use cases, since AI systems share lifecycle processes [ISO23c]. The, in the next chapter 4 proposed comprehensive and flexible QG-based IM structure for concrete design decision-making aims to mirror and monitor the overall lifecycle concept. This includes how design decision-specific information relates with relevant contextual information.

Bridging the gap from RAI knowledge generation to RAI knowledge application through IM, we identified two MQG4AI-interaction scenarios, which are envisioned to enable a decentralized and on-going approach: public MQG4DesignKnowledge, and private MQG4Application that are able to communicate with one another through a shared IM structure. MQG4A provides project teams with a shared lifecycle conceptualization template that is filled with concrete guidelines for continuous planning of individual AI lifecycles based on MQG4AI, including (future) use case-adapted MQG4DK blueprint contributions that are based on (novel) RAI design knowledge, possibly derived from MQG4A scenarios. These two envisioned MQG4AI interaction perspectives are further outlined in the following sections, before introducing the design process to build the basic MQG4AI information structure in chapter 4.

3.3.7.2. Public MQG for Design Knowledge (MQG4DK)

The MQG4DK living lifecycle blueprint provides the foundation to align RAI design knowledge through a generic and customizable information structure for AI lifecycle planning, i.e. MQG4AI. We focus on the ever evolving and dynamic AI lifecycle design and risk knowledge.

Intention Overall, we aim to enable the creation of a growing RAI ontology to organize generalizable instructions for the domain- and use case-adapted implementation. We envision to facilitate the continuous organization of RAI lifecycle design knowledge for application by providers and regulators alike. In addition, MQG4DK fosters RAI literacy at the interface of regulation and implementation through providing access to qualitative guidelines. The envisioned MQG4DK blueprint is designed to evolve in a decentralized manner, responding to AI's evolutionary character, as well as the scattered state of RAI design knowledge and a dynamic real world. We intend to design the foundation for further contributions based on the MQG4AI RAI information structure, as proposed in this thesis.

Quality through Unified Information Processing The desired quality of accepted methods is achievable based on structured processing and transformation of lifecycle design knowledge that is linked to relevant AI QMS [Fut24] information in form of QGs along a generic and use case-specific AI lifecycle flow. The, from existing or novel contributions extracted guidelines for AI implementation design are organized along the lifecycle, and linked with AI system-specific information, as well as identified AI risks in form of QGs. As a quality-check, QGs focus on the inclusion of practical guidelines, which comprises the foundation for MQG4AI's proposed lifecycle design. As introduced in section 1.3.1.2, we intend to design MQG4AI in alignment with criteria for acceptable standards for high-risk applications, as communicated by the EU AI Office [SG24b] [Eura]. Thus, the design of individual QGs that mirror concrete design decisions, and are mapped along the generic AI lifecycle, aims to include clear, application-oriented instructions and qualitative methods, enabled by the proposed generic and customizable information structure, as presented in the next chapter 4. Knowledge contributions include contextual information, such as AI risks, as well as lifecycle design decisions that are linked and the, in section 4.3.2 outlined *QG-Naming* structure is intended to organize decentralized contributions according to shared structural similarities of AI techniques.

3.3.7.3. Private MQG for Application (MQG4A)

MQG4A is intended to provide a tool for application during individual AI projects for continuous lifecycle planning through IM that enables the inclusion of RAI design knowledge.

Intention As promoted by the European AI Office [SG24b] [Eura], MQG4A is envisioned to be applied as an additional management layer during individual AI projects. Focusing on contributing stakeholders along the AI lifecycle, from the RAI design knowledge derived templates provide a method for continuous AI information-based lifecycle conceptualization.

Lifecycle Evolutions During project application, different MQG4A-versions reflect the evolution of individual lifecycles for planning, and, in contrast to MQG4DK, they include concrete results. For instance, focusing on model development, MQG4A-versions are envisioned to structure the iterative process of reliable lifecycle design decision-making as part of an iterative *conceptualization* phase. We propose to structure design decision-making into pre-, intra-, and post-selection versions [EM24] for reliable lifecycle planning, as illustrated in chapter 8. When making a model development-related design decision, the process begins with identifying prerequisites, such as data composition. Next, related design choices are appended to provide further clarity, and finally, optimization techniques are implemented to enhance the outcomes, as further detailed in chapter 2. This generic structure is intended to be generalizable to establish relevant design decisions, such as reliable performance metrics on which decisions for model optimization are based, focusing on interdependency extraction and iterative design decision making. Generally, model *optimization* is based on the comparison of evolving results with modifications to a baseline approach until reasonable updates are identified, and a reliable model development concept established. We aim to mirror this strategy in form of MQG4A template versions to log results with lifecycle concepts. Other generic lifecycle stages might need a different approach. The envisioned interplay between MQG4A and MQG4DK is outlined next.

3.3.7.4. Envisioned MQG4DK & MQG4A Interplay

With respect to MQG4DK, MQG4A applies/provides/updates/transforms RAI knowledge, adapted to the use case at hand in an optimized form compared with other MQG4DK-contributions through a shared MQG4AI information structure by default. Figure 3.16 illustrates the envisioned interplay between these lifecycle blueprint versions from a conceptual perspective. Public contributions to MQG4DK, as well as private application of MQG4A are intended to continue MQG4AI's four fundamental design principles, see section 3.3.2, through the, in the present thesis identified lifecycle blueprint information structure towards RAI by design.

The use case-adaptation of MQG4A is envisioned to be realized by a configurable pull version from MQG4DK (MQG4A-v0). This dynamic is carried out by MQG4AI's fundamental building blocks: QGs, and their proposed information processing format are introduced in the next chapter 4 in more detail. Therefore, relevant QG information layers, i.e. *QG Naming* and *QG Tags* are outlined in section 4.3.2. They are intended to enable an intelligent search through attributes that specify each appended design decision's applicability from a methodological viewpoint. As a result, only applicable lifecycle design decisions are pulled from the design knowledge base. Simultaneously, they enable RAI design knowledge

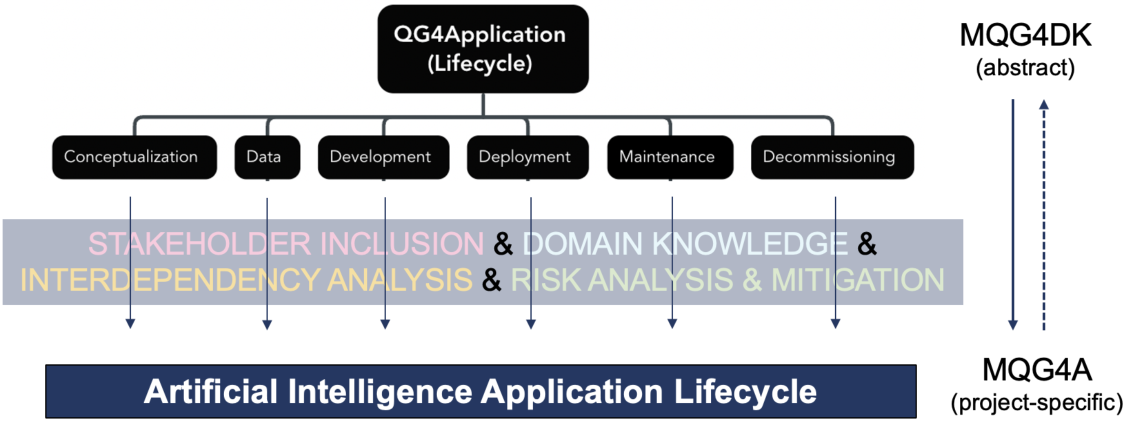

QG4Application (Lifecycle)
Conceptualization
Data
Development
Deployment
Maintenance
Decommissioning
STAKEHOLDER INCLUSION & DOMAIN KNOWLEDGE &
INTERDEPENDENCY ANALYSIS & RISK ANALYSIS & MITIGATION
Artificial Intelligence Application Lifecycle
MQG4DK
(abstract)
MQG4A
(project-specific)

Figure 3.16.: Information interplay between MQG4DK and MQG4A. The root node *QG4Application* summarizes information on all underlying QGs, and QG-Scoring is introduced in section 4.1.3. For example, this leads to varying AI lifecycle evaluation scenarios from the perspectives of MQG4DK and MQG4A. The former offers information on potential evaluation setups across different structural contexts, organized by QG-naming and QG-tags (see section 4.3.2), and provides corresponding design guidelines. The latter delivers concrete outcomes for specific AI projects, enabling the setup of a comprehensive scoring system that considers all connected lifecycle QGs.

organization within MQG4DK. Finally, the generic default information for high-risk systems that defines the basic MQG4AI lifecycle blueprint, as introduced in this thesis closes the gap between MQG4A and MQG4DK.

In summary, this section describes our method *Design Science Research*, based on which the proposed dynamic between MQG4A and MQG4DK is envisioned to continue in a decentralized manner answering to AI's evolutionary character and a constantly changing world. This is clarified through the extraction of generalizable workflows to fill and utilize the blueprint in part III. First, the next section lays out our vision and limitations of our proposition, before introducing MQG4AI's building blocks in more detail in the next chapter 4.

3.4. Vision & Limitations

After introducing our scope and method for MQG4AI design in section 3.3, as well as its context RAI in section 3.1, with a particular emphasis on RM in section 3.2, we outline our vision and limitations regarding the MQG4AI lifecycle blueprint in this section.

Since the outlined project is an immense endeavor to complete, we envision a decentralized continuation of our proposed approach to contribute to AI QMS. Overall, the proposed foundational MQG4AI RAI information structure needs to be elaborated further, and continually refined, as well as tested for concrete application scenarios, and the continuation of customizable information blocks. Further, the conceptual setup, as outlined in the next chapter 4 in more detail, will need to be tested and refined with more use cases and on-going decentralized knowledge contributions from a broad pool of experts. We envision to enable continuous and decentralized MQG4AI interaction. Therefore, we lay the groundwork for further contributions based on the extracted blueprint design workflow in section 3.3.7, which is concretized in part III to continue MQG4AI design. There, we illustrate how to apply the identified building blocks, so that decentralized knowledge contributions to MQG4DK, as well as MQG4A-scenarios are possible. Finally, to optimize the MQG4AI roadmap, which is explored in chapter 9, we propose to implement a first lifecycle prototype with the support of GenAI. Aiming to parallelize MQG4AI design and testing, we intend to design individual building blocks in a flexible and interactive manner.

In this section, we begin with outlining how MQG4AI aligns with AI Act-conform AI QMS requirements [Fut24], that provide guidance for required RAI information collection, before introducing next steps to refine the blueprint in more detail.

3.4.1. Customizability towards AI Quality Management

AI QMS, as outlined by the AI Act in Article 17 [Fut24] comprise the foundation of required information to implement RAI at the interface of regulation and implementation. We aim to develop a concept that contributes to AI QMS through IM, which addresses the technology's underlying dynamics. The proposed generic and customizable design is envisioned to allow for continuous knowledge updates, so that it adapts to AI's evolutionary character within various contexts. First, we broadly introduce MQG4AI design, next we outline how it corresponds to AI QMS, before highlighting open issues in the following sections.

3.4.1.1. Flexible MQG4AI Design

We design the MQG4AI lifecycle blueprint through a retrospective analysis of use cases and research on RAI, following DSR [vJ20], as introduced in section 3.3.7. We envision to enable flexible docking of contextual information to the lifecycle to enable bidirectional communication, following the identified IM structure that focuses on interdependencies and iterative cycles. Starting with AI system-specific and AI risk-related information that is linked with the lifecycle concept through QG, we aim to introduce customizability to incorporate additional, contextually relevant information blocks that contribute to AI QM.

Central MQG4AI Building Block We outline MQG4AI's central building block *Quality Gate* in the next chapter 4 in more detail. Broadly, two types of QGs exist: *Collection-QGs* that construct the AI lifecycle, and *Leaf-QGs* that mirror concrete design decisions. Based on a retrospective analysis of our experiments, we derive the customizable *leaf-QG*s to process design decisions along the AI lifecycle. We aim for a reliable information format addressing the technology's inherent dynamics towards quality by design through a unifying information structure. This results in our proposed customizable *leaf-QG template*. It comprises a combination of information layers and allows to append additional layers, as detailed in section 4.3. The template equally provides information layers based on which we illustrate how to append reliable design decisions to MQG4DK and pull MQG4A-versions.

3.4.1.2. MQG4AI & AI QMS Requirements

The proposed MQG4AI information and building blocks are envisioned to provide flexibility regarding their definition and arrangement. With this setup we aim to enable information integration with existing QM procedures outside the MQG4AI blueprint that already organize relevant information, for instance based on existing sector-specific regulation. The, in section 3.3.3.3, outlined EU AI Act-conform documentation requirements complement the argumentation provided in this section. MQG4AI, in its current form, incorporates AI QMS requirements outlined in Article 17 of the EU AI Act [Fut24], and we are confident that the proposed setup is adaptable to encompass information on all related requirements.

- (1.g) Risk Management System (refer to Article 9)

The inclusion of a RMS is at the center of MQG4AI and was previously introduced. Addressing the need for continuity and iterations of identified RM processes and risks is outlined in the next section 3.4.2.1, analogously to iterative lifecycle design decision-making.

- (1.f) Data Management

- (1.h) Post-market Monitoring System (refer to Article 72)

MQG4AI's generic lifecycle structure provides room to include information on the data lifecycle, which we superficially design within our contribution in section 4.2, while generally considering model development-related interdependencies. For instance, focusing on the development phase, leaf-QGs provide the option to extract post-market monitoring information from individual design decisions. With respect to applicability during maintenance the addition of a post-market monitoring module, comparably to the RM information block, might be reasonable to survey the extracted information from leaf-QGs, as detailed in section 4.3, combined along the AI lifecycle. This information could equally be summarized in form of a scoring index within the root-QG, which is introduced in section 4.1, along other relevant concepts.

- (1.a) Strategy for Regulatory Compliance
- (1.b) Design, Design Control, Design Verification
- (1.c) Development, Quality Control and Quality Assurance
- (1.d) Continuous Examination, Test and Validation procedures

MQG4AI addresses the interface between regulation and creation. It is envisioned to contribute to these four requirements by design, which needs to be tested and evaluated for concrete high-risk AI projects that utilize the blueprint for lifecycle planning as part of future work, necessitates transformaing MQG4AI into a usable tool, first.

- (1.e) Technical Specifications
- (1.i) Reporting of serious incidents (refer to Article 73)
- (1.j) Communication with National Competent Authorities
- (1.k) Record Keeping
- (1.l) Resource Management
- (1.m) Accountability Framework

While we focus on RM and related system information structured along the AI lifecycle, other QMS-related information can be included and linked in a similar manner. Also, leaf-QGs' customizability through the option to append additional information layers results in flexibility of information extraction based on lifecycle design decisions, for instance regarding resource-related computing power. Comprehensive resource management might necessitate an additional information block, depending on the respective scope and more empirical experience with MQG4AI interaction. Possibly, it might be more reasonable to append this information as a submodule to the already existent system information block, and/or summarize it as a scoring-index within the root-QG.

In summary, we envision to design a multi-functional information structure for application, and continuous AI design KM. In addition to supporting AI QM, this approach is intended to enable analysis that foster our understanding of AI as a technology towards broad AI literacy. For instance, through analyzing comprehensive information on design decisions, such as performance metrics based on where and how they appear along the AI lifecycle across multiple use cases. There is still much to be done. The following sections outline possible next steps in designing the MQG4AI lifecycle blueprint, building on our contribution as a starting point for a broader roadmap.

3.4.2. Completing & Testing the MQG4AI Lifecycle Blueprint

We begin designing MQG4AI focusing on RAI-related IM, which comprises knowledge on TAI and AI risks, focusing on reliable design decision-making during model development to illustrate how to fill MQG4DK with evolving RAI knowledge, as well as envisioned MQG4A utilization during individual AI projects. The extracted layout is derived based on experiences during development, other stages possibly necessitate a different approach. Therefore, identified information collection and building blocks that provide the mechanisms for a connected information flow need to be tested for applicability along the AI lifecycle. The snapshot, as outlined in the present thesis, is envisioned to enable completing and refining the living blueprint in a decentralized manner through concept definition and workflow extraction, Overall, MQG4AI, if adopted would benefit of knowledge from a diverse pool of experts world wide. This section summarizes our contribution and highlights possible next steps, aiming to transform the blueprint into an applicable tool to enable interaction-scenarios.

3.4.2.1. *Quality Gates*, AI Risks & AI Lifecycle Phases

Our main contribution comprises laying the foundation for the MQG4AI concept. We propose a first approach to implement a lifecycle planning blueprint that incorporates supplementary information blocks, as well as offers a unifying information structure (collection-, and leaf-QGs) that incorporates RAI-specific interdependencies. These components for IM towards quality by design are intended to provide sufficient generalizability across domains and AI scenarios, including a fluid and adjustable interpretation of *Design Decisions*. However, possibly other realizations of the MQG4AI components are necessary for different lifecycle phases and application scenarios. Overall, we derive the current version of MQG4AI from experiences during model development, focusing on reliable information linking between AI lifecycle design decisions and with contextual RAI information. AI is currently a highly discussed topic, with numerous standards and guidelines being published. Intelligent systems and their implementation are a living process that continuously evolves. Consequently, based on more use cases, including other domains, and lifecycle stages, the blueprint's first draft can be refined. This continuous feedback process is implemented through MQG4AI's method, the DSR process, as introduced in section 3.3.7.1.

Generic AI Risk Information Block We organize risks based on TAI [Hig20], establishing a foundation for generalizable AI risk classification, and aiming for transferability to individual project planning, which includes relevant system-specific information. With the proposed customizable setup, we envision connectivity of these contextual Information Blocks with the AI lifecycle, as well as interchangeability with external, already existing resources and information collections. Our goal is to enable decentralized risk identification, as well as linking of controls and risk sources with the lifecycle implementation as first steps towards embedded RM, closing the gap between generalizable and use case-specific guidelines. Information linking with the AI lifecycle is enabled by QGs, the central building block of MQG4AI. We simulate these mechanisms for several risks, including a selection of risk controls in part III, associated with the model development lifecycle stage. The extracted MQG4AI interaction workflows require further extension and testing, e.g. encompassing additional RM procedures within individual projects by design.

Generic AI Lifecycle We establish six generic AI lifecycle phases, with a detailed exploration of the *Development* stage based on our use cases, emphasizing information flow of interdependent design decisions. MQG4AI is envisioned to function as a tool to continuously execute the lifecycle design, or *Conceptualization* phase. To create an application-oriented lifecycle, each phase is refined through practical experience, necessitating a lifecycle creation process that is both abstract and adaptable to diverse application scenarios. This degree of flexibility is enabled through collection- and leaf-QGs, the foundational MQG4AI building blocks that equally support the integration of contextual RAI information. Our contribution covers the identification of generalizable sub-processes of parts of model development, focusing on model *Evaluation* and *Explanation*, as well as related *Data* utilization for model-related *Preprocessing*. We include information extraction for other lifecycle stages, as well as linking with supplementary contextual information focusing on AI risks. The other stages, highlighting *Data*, *Deployment*, *Maintenance*, and *Decommissioning*, need to be defined in more detail through practical expertise, possibly following the, in this thesis extracted information linking approach. Finally, the proposed skeleton to organize high-level lifecycle stages, which is detailed in section 4.2, needs to be further approved through application to more use cases.

Leaf-QG Information Layer The proposed information processing template for lifecycle design decision-making, which is based on experiences during development needs to be further refined and tested, aiming for practicality during lifecycle phases beyond development, and supporting the advancement of AI's on-boarding in real world contexts. This includes the identification of optional and required information layers, aiming for quality by design of leaf-QG contributions to MQG4DK, as well as testing more additional information extraction layers in form of MQG4A. Our contribution exemplifies related risks, as well as the extraction of information for post-market monitoring, as outlined in section 4.3, while the application of this information during later stages needs to be further analyzed.

3.4.2.2. Multi-User Interface & MQG4AI Software

As part of our contribution, we design the global MQG4AI concept, and once sufficiently approved and refined, the envisioned interaction-scenarios with the MQG4AI lifecycle blueprint need to be implemented as a software to enable decentralized applicability.

MQG4AI Architecture For an overall smooth application, the creation of MQG4AI's customizable information and building blocks (collection- and lead-QGs) needs to be implemented, as well as bidirectional information linking automated. A suitable software architecture, enabling MQG4DK and MQG4A scenarios, see section 3.3.7, needs to be identified. This includes the implementation of quality checks that are enabled through the information processing structure by design. For instance, if a non-optional *Leaf-QG* layer, such as *Input Information* is left empty, this needs to be handled accordingly. Lifecycle design decisions are interconnected and require additional information. Thus, customizability is required to enable a decentralized modification of information layers and blocks.

User Interface The user interface should adapt to different interaction scenarios. Aiming for usability, we highlight the intersection of regulation and implementation. It is characterized by a multitude of stakeholder roles, as outlined in sections 3.3.2.1 and 3.3.3.5, which results in a possibly differing information design for stakeholders that need to consult the lifecycle implementation from different perspectives. In addition, stakeholders that contribute to MQG4DK require an interface for design knowledge contributions.

MQG4A & MQG4DK The implementations of MQG4A and MQG4DK within MQG4AI differ significantly and are highly dependent on the chosen system architecture. For example, in the context of compliance assessment, MQG4A requires mechanisms to generate documentation from information extracted during specific AI projects, a functionality not needed for MQG4DK. Additionally, an intelligent search feature must be developed to retrieve configurable MQG4A variants based on leaf-QG-tags, and a hosting solution is required for MQG4DK to ensure decentralized accessibility and integration.

3.4.3. MQG4AI Interaction & Continuous Refinement

Once the fundamental MQG4AI lifecycle blueprint is sufficiently finalized and a first prototype implemented, concrete MQG4A-scenarios can be tested, and decentralized MQG4DK RAI knowledge contributions are enabled. This section outlines envisioned workflows for MQG4DK and MQG4A interactions, highlighting the organization of RAI knowledge contributions to MQG4DK, and pulling of MQG4A-versions.

3.4.3.1. Workflow Extraction for MQG4DK Design

Based on the generic MQG4AI-structure, we start filling the MQG4DK living blueprint with generalizable RAI information for a selection of AI risks and risk mitigating design decisions (leaf-QGs) along the AI lifecycle (collection-QGs). We aim to extract workflows that are envisioned to inspire decentralized information contributions. The proposed generic collection-QGs that construct the AI lifecycle need to provide sufficient generalizability to combine diverse RAI knowledge contributions in form of leaf-QGs. We extract two different workflows to illustrate and design MQG4DK contributions, as further detailed in part III.

Lifecycle Design A possible workflow towards designing a risk-mitigated AI lifecycle is outlined in chapter 6, where we align AI risks and best practices focusing on the *Transparency*-related criterion *Explainability* of intelligent systems [Hig20], emphasizing XAI methods. Based on those findings, we construct the generic lifecycle explanation stage during model development, and illustrate a leaf-QG-contribution to MQG4DK for a technical guideline to evaluate the quality of LIME and SHAP explanations.

Complex Design Decisions As an additional contribution scenario, chapter 7, outlines MQG4DK-contributions for a combination of related design decisions that address the risk *Unreliable Performance Evaluation Metrics*. Based on those findings we highlight information linking with contextual system information, and further design the *Evaluation* stage during model development, which provides a solid basis for MQG4A scenarios. We additionally illustrate linking with system-specific information, and our contribution is centered around a fictional use case, as introduced in section 5.2.1.

Future Workflows Other workflows for MQG4DK contributions are possible, while we highlight the importance of linking AI lifecycle design decisions with related risks and other contextual information to foster the desired quality of acceptable design decisions. This includes linking with AI system information for MQG4A, and, if applicable abstracted to MQG4DK, for instance for MQG4A-based contributions. This interplay needs to be tested for individual use cases that apply MQG4A during AI projects for lifecycle planning, and simultaneously contribute RAI knowledge based on the individual use case to MQG4DK.

3.4.3.2. MQG4A Scenarios

Concrete MQG4A scenarios are part of future work and need to be tested. We envision to enable iterative MQG4A lifecycle planning scenarios based on the generic high-risk MQG4AI information structure. It is continued with decentralized MQG4DK contributions, and guidelines are pulled to MQG4A, if applicable to the respective use case, which is implemented through leaf-QG-tags and QG-naming. Following DSR [vJ20], the MQG4AI concept can be refined through application to specific (medical) AI research projects (in form of MQG4A) addressing questions surrounding practicality of the design, as outlined in the previous sections until a sufficiently refined foundational blueprint version is identified.

MQG4A Workflow The proposed structure to organize derived template versions during application into pre-, intra-, and post-selection steps was empirically extracted based on experiments during model development for design decision-making. Concretely, we extracted the selection method for reliable performance evaluation metrics of multi-label ECG classification [EM24], while the process is intended to be generalizable to other lifecycle design decision-making, such as the selection of the loss function or model architecture. This depends on the required degree of complexity and setup of individual *Design Decisions* along the AI lifecycle. In contrast to MQG4DK, MQG4A leaf-QGs include concrete results that enable an iterative evaluation of design choices. This can be reflected as a combination of MQG4A template versions for individual design decision-making. We simulate a possible MQG4A scenario in form of a retrospective analysis to outline how we envision template-versions to be applied for design decision-making as part of a concrete use case in chapter 8. The segmentation model configuration process is embedded within a medical software that enhances digital support of the rare disease of the esophagus *Achalasia*, as detailed in section 5.2.2.

3.4.4. Parallelization

As outlined, there are still many open questions before MQG4AI can be fully utilized. Overall, we see a significant need for concrete tools that enable AI QM and assist in developing RAI at the intersection of regulation and the creation of intelligent systems. Through this work, we aim to contribute an idea that could be implemented as a tool. We explain the basic concept as comprehensively as possible in this thesis. Thus, we intend to ensure the core idea of how MQG4AI is envisioned to provide support for AI QM through continuous IM, is clear, should it gain traction. Overall, our contribution results in a *living blueprint* that continuously evolves in alignment with dynamic RAI knowledge. For implementation, it is possible to parallelize certain steps. For example, refining lifecycle design and the development of an initial software could proceed simultaneously, especially with the support of GenAI. As a result, testing the application for an initial use case would be feasible. However, this would require a larger, multidisciplinary research team and a long-term research project to work out the finer details.

In summary, this chapter presents MQG4AI, its context centered around AI QM, while highlighting RM towards implementing RAI systems. This results in MQG4AI's information blocks that constitute the basic lifecycle blueprint, namely, RM and AI system-relaed information. Finally, envisioned interaction scenarios to populate the blueprint with further RAI knowledge are outlined based on DSR [vJ20]. Finally, the outlined vision and identified limitations are continued in chapter 9, summarizing the MQG4AI roadmap and research objectives, in case the concept achieves recognition. QGs comprise the foundational building block to enable the envisioned comprehensive RAI lifecycle planning blueprint. The following chapter 4 describes the information processing concept of MQG4AI to design the AI lifecycle in more detail, aiming to close the gap from generic to use case-specific implementations under the inclusion of RAI design knowledge.

4 Setup, Components and Structure

This chapter delves into MQG4AI's proposed conceptual setup and building blocks that construct the AI lifecycle to realize the, in the previous chapter 3, introduced vision towards implementing RAI. Overall, the MQG4AI lifecycle blueprint, which is illustrated on GitHub[1], aims to provide a unified information structure for comprehensive lifecycle planning to *Contributing Stakeholders* with shared access. It is based on the premise that every lifecycle design decision contributes to achieving the desired quality of AI systems. The methodology's components are designed to be both generic and customizable, addressing the dynamic nature of AI while integrating design decisions with relevant contextual information. The previously, in section 3.3.4, broached AI lifecycle information block is at the heart of our proposed contribution to AI QM through RAI knowledge-based IM, is detailed in the following sections.

We introduce the two main types of *Quality Gate* (QG) (collection- and leaf-QGs) that reflect RAI lifecycles in section 4.1. These elements form the core of the blueprint, guiding the lifecycle from generalizable to use-case-specific implementation guidelines, under the consideration of contextual *RAI Design Knowledge*. Next, we outline the generic lifecycle that aims to provide sufficient flexibility to enable the fusion of design concepts across use cases in section 4.2. The leaf-QG template for information processing, which is envisioned to enable high quality by design through guided and reliable design decision-making, is described in section 4.3. Finally, we introduce challenges regarding AI in medicine, as well as two medical use cases in the next chapter 5, based on which we evaluate MQG4AI interaction scenarios in part III.

4.1. Concept *Quality Gate* for RAI

Generally, the concept *Quality Gate* (QG) finds application across various domains that need to ensure high-quality standards, and follow a creation process that can be organized into phases. For instance, the authors of [FM20] define elements

[1] `https://github.com/miriamelia/MQG4AI/blob/main/README.md`

of QGs in the context of manufacturing, including the abstraction of a *morphological box* that enables the identification of manufacturing QGs for application to individual use cases. We highlight QGs as approached in software engineering and product development settings, that are centered around stages along a system lifecycle, as well as adhere to user requirements, resulting in similarities with AI systems, as introduced in section 3.3.4. On an abstract level, QGs are integrated within a process, perform an (iterative) evaluation of identified desirable criteria that, if fulfilled allow the process to pass on, otherwise changes of the executed procedure need to be conducted. Consequently, QG realization is centered around the use case-adapted definition of criteria, as well as some form of evaluation, e.g. thresholds or metrics that assess criteria to decide on sufficient fulfillment for passing, in addition to positioning QGs along the use case-specific process chain. [GM09] [HS09] [Fil06] [Flo08]

In contrast to traditional software, intelligent systems are non-deterministic, and AI's opacity, as well as evolutionary character, and use case-specificity result in the need for a corresponding approach to implementation and design, continuously ensuring the desired behavior in the real world. As a result, we promote the idea of quality through *RAI Information Management* (IM), focusing on comprehensive and continuous AI lifecycle planning. Our viewpoint to kick-start MQG4AI is through the lens of developers. At its core, this setup is realized through QGs, and we demonstrate the AI-adapted QG information structure towards RAI by design in the following sections.

4.1.1. QG Interdependency Graph, Positioning & Scope

First, QGs need to be identified and positioned in a reasonable manner in alignment with the specific use case, creating a QG layout. Overall, QGs "[...] structure a process chain into phases and allow a periodical review of the process quality" [HS09, 206], mirroring a significant milestone or decision point within a project [Flo08, 245]. Consequently, a process, consisting of a compilation of QGs is constructed [Flo08, 250], that reflects the individual project's underlying stages. This entails the verification of QG positions, including flexibility towards optimization of the identified structure [HS09, 211]. Also, some identified QGs might be optional or weighted differently. This section introduces QGs in the context of AI, which are characterized by interdependencies, and positioned along the AI lifecycle, shaping their scope on information access.

QG Interdependency Graph All processes and design decisions that result in the intelligent system are organized along the AI lifecycle, which is introduced in section 3.1.3. Its realization is dynamic, covering generic to use case-specific phases and design decisions. Therefore, QGs are arranged in a tree-structure starting from a generic high-level lifecycle that is shared between use cases, which we commence to design in section 4.2, and leading to more flexibility in QG-identification on lower levels. The resulting bidirectional *QG Interdependency Graph* is based on two types of QG (collection- and leaf-QG) that function as basic building blocks to implement IM, focusing on vertical and horizontal information extraction and linking, as well as reliable information processing of concrete design decisions. Their design is introduced in more detail in the next sections.

QG Positioning In the context of MQG4AI, generic and customizable QGs are positioned along the AI lifecycle, aiming for RAI IM. *Collection-QG*s construct vertical levels of AI lifecycle stages in a flexible way and are derived based on the identification of generalizable workflows for lifecycle stage execution. They are hierarchically structured, and specify the, in section 4.2 outlined generic level of the AI lifecycle to summarize and organize related design steps, aiming for planning and documenting qualitative lifecycle implementations. This assignment of lifecycle QGs constitutes *QG Positioning*. For instance, the *Development* stage is organized into the generalizable interrelated collection-QGs model *Configuration*, *Evaluation*, *Optimization*, and *Explanation*, that each comprise lower levels. *Performance Evaluation* in turn is defined by sub-processes such as the identification of a suitable *Metrics Compilation*, including *Additional Material* for accurate and use case-adapted interpretation, including the definition of performance evaluation *Objectives*. This may result in a compilation of leaf-QGs for individual design decisions at lower levels, which we demonstrate in chapter 7.

QG Scope Overall, collection-QGs identify vertical interdependencies along the AI lifecycle, and different stages can include up to n sub-collection-QGs. Each *QG Scope* is defined in alignment with its hierarchical level. In addition to vertical levels, the scope includes horizontal interdependencies that identify relevant project-based resources[2] or outcomes of other QGs, that influence or are influenced by the respective QG. The scope is identifiable on all levels, while higher level collection-QGs are defined by the combined horizontal interdependencies of lower levels, which is extracted from the leaf-QG information structure at the lowest level of the QG-graph. For instance, IM is achieved through the *leaf-QG* format, which reflects concrete design decisions and processes information about their definition, including relevant horizontal input and output information, and the transformation of relevant information tailored to different stakeholders. The customizable assignment of leaf-QGs to collection-QGs is intended to bridge the gap to use case-specificity in form of structured information on concrete implementations guidelines. Leaf-QGs, or compilations thereof mirror concrete *Design Decisions* tailored to specific AI techniques and are detailed in section 4.3.

Finally, *QG Relations* may emerge that characterize how individual QGs relate to one another. We propose to define relations based on object-oriented programming (OOP) [Par20] and explore *QG Inheritance* in chapter 7 for MQG4DK.

[2] The lifecycle implementation is shaped by contextual information, which MQG4AI illustrates for RM and information on the system's real-world setting, as outlined in section 3.3.

4.1.2. QG Application, Criteria & Gate-keeping

In correspondence with the identified QG layout, that typically positions QGs at transition points between a project's phases, iterations, or increments [Flo08, 248], QG application entails "[...] selecting and evaluating the correct criteria at each gate and, applying appropriate rules to gate decisions [...]" [GM09, 202]. These criteria enable quality assessment by an instance other than the developer, when the system passes from one state to another state [Fil06]. This process is called gate-keeping, and refers to the "[...] decision whether a project may proceed or not" [Flo08, 245], which comprises a component of QM [Flo08, 245]. Generally, QGs can be employed as either a *quality guideline* or a flexible *quality strategy* in project QM, which impacts their criteria selection. As a *quality guideline*, they establish a consistent set of criteria applied across projects to ensure a comparable minimum quality level. Alternatively, as a *quality strategy*, they are adapted to meet the specific needs of individual projects, offering flexibility, while maintaining focus on project-specific objectives. [Flo08, 245] In addition to a QG application strategy, the question how to define criteria for QGs should address continuous updates to ensure their on-going quality, and a multi-stakeholder approach for criteria evaluation, or gate review, is recommendable [Flo08, 251]. Overall, criteria for QGs are most effective when defined systematically by deriving them from abstract (business) goals. Methods such as Goal-Question-Metrics (GQM) or quality models, as introduced in [KV22] for AI, can be employed for this purpose, though they are often avoided due to their labor-intensive nature. Instead, companies frequently rely on the expertise of senior developers, an experience-based approach that may not ensure full alignment with objectives or consistently deliver robust criteria. [Flo08, 248]

QG Application In the context of MQG4AI, we follow a decentralized approach, as introduced in section 3.3.7, resulting in different perspectives on required information density depending on MQG4A or MQG4DK interaction scenarios, which we only broach for QG concept introduction in this section, as they represent two sides of the same coin. This is clarified in more detail through the illustrated MQG4AI interaction scenarios in part III. Overall, we aim for quality by design through iterative IM that incorporates human design knowledge, and provides structured information processing and linking to different stakeholders. Therefore, *QG Application* combines elements of a consistent *quality guideline* and a flexible *quality strategy* [Flo08, 245]. The former is defined by the generic AI lifecycle, which provides a dynamic blueprint that consists of required process steps and design decisions, closing the gap to use case-specificity on lower levels, as previously introduced. In addition, the leaf-QG template provides a generalizable information structure to capture stakeholder knowledge in a unifying way. Flexibility is provided on lower levels of the QG graph. QG positioning happens in alignment with individual projects for MQG4A and conceptual implementation graphs my differ, and MQG4DK contributions are made with respect to specific AI techniques along the generic structure. Further, MQG4DK contributions require a comprehensive approach to information density aiming to convey the desired quality of acceptable methods. In contrast, regarding MQG4A-scenarios, the proposed leaf-QG information layers are filled optionally, if reasonable, depending on the respective design decision. This addresses real world projects, where the lifecycle is defined iteratively. Especially, during development, information on the lifecycle implementation grows gradually, which results in different lifecycle versions with different levels of information density. Overall, this dual approach allows for balancing standardization with adaptability, aiming for quality assurance through generic and customizable AI lifecycle planning and RAI knowledge organization.

QG Criteria Aiming for a RAI lifecycle implementation, *QG Criteria* comprise if and how lifecycle process stages and design decisions are executed. In addition to collection-QGs along the AI lifecycle, we design individual leaf-QGs as a template to organize RAI information, focusing on responsible design decision-making. Leaf-QGs mirror design decisions and link horizontal information, continuing the identification of interdependencies with respect to the vertical collection-QG structure that results in leaf-QGs. Our starting point for defining the leaf-QG template focuses on continuous conceptualization of parts of the lifecycle *Development* stage, broaching the strongly related *Data* stage, and including information extraction for *Deployment*, as well as *Maintenance*. The proposed information processing template, as detailed in section 4.3, may need to be adjusted for other lifecycle stages and AI system information. Overall, the unifying structure to document lifecycle design decisions aims to direct stakeholders towards risk mitigation and quality by design through IM of implemented lifecycle concepts, including contextual information, and under the consideration of *Guiding Questions* [IA12], in response to the AI Act's risk-based approach. Finally, the provision of ethical and domain-specific information within MQG4AI (see section 3.3.3), as well as the option to extend RAI information blocks, is equally intended to contribute to a high-quality implementation. Possibly, the provision of a reward system could motivate stakeholders to access supplementary content surrounding RAI for information linking within their individual lifecycle scope, in addition to project-wide events, as further considered in section 3.2.2.3.

QG Gate-keeping QGs are evaluated in a decentralized manner though identifying how the leaf-QG layers are filled,[3] as well as which collection-QGs exist for MQG4A scenarios, resulting in *QG Gate-keeping*. This equally comprises considered contextual information blocks. Leaf-QGs are defined by an information template that is intended to provide necessary channels ensuring the desired quality through customizable information processing and linking, which is envisioned to result in automatic testing of design decision-making. At their core, we propose to guide *Leaf-QG* creation through a three-fold structure (*Content* definition,

[3] As previously outlined, MQG4DK requires an information-dense approach, while MQG4A scenarios may result in less information depending on the project's evolutionary state.

Method extraction, and content *Representation* tailored to stakeholder-views) for reliable design decision-making, including an *Evaluation* layer to shed light on possible open questions regarding the chosen method. Highlighting the identification of relevant *input and output information*, we aim to enable a comprehensive interdependency analysis, which is crucial to identify, and avoid AI pitfalls by design. The leaf-QG's layered structure is envisioned to contribute to the implementation of reliable lifecycle concepts, as well as to provide information extraction for additional AI QMS requirements such as post-market monitoring, which is influenced by the individual design concept. Further customizability is provided through the option to remove/append additional related information layers, which we illustrate for the *risk management layer*, aiming to integrate related information on AI risks with concrete design decisions. Finally, leaf-QGs play a crucial role in enabling MQG4AI's decentralized, and continuous utilization, which paves the way for on-going quality updates. Possibly, future work contributes to the design of more global lifecycle blueprint evaluation mechanisms.

4.1.3. QG Information Evaluation & Scoring System

Our proposition how to transfer QGs to the context of AI through IM and KM towards QM, aligns with [HS09], who extend "[...] the Quality Gate approach by methods for identifying critical information flows [...]" [HS09, 211]. They incorporate "[...] dynamic, information creating and processing" [HS09, 207] to enhance product development, highlighting flexible positioning of QGs [HS09, 209] within processes. With respect to the generic AI lifecycle blueprint that contributes a starting point how to organize QGs within MQG4AI's lifecycle information block, our lifecycle planning methodology exhibits similarities with *roadmapping*: "[...] roadmaps are guiding structures that facilitate the execution of a set of pre-defined high-level activities that have been deemed necessary to achieve specific objectives" [GM09, 201]. The authors of [GM09] equally promote IM along the product (AI) lifecycle, namely "[...] roadmaps serve as a means of structuring and managing the information and explicit knowledge [...]" [GM09, 201]. In addition, the integration of QGs with *roadmapping* is deemed beneficial, supporting reasonable and efficient QG positioning, as well as QG criteria selection, and gate-keeping procedures [GM09, 202]. Further, addressing the quality of QG criteria, which in the context of MQG4AI are designed towards risk mitigation, they equally highlight the role of addressing RM during project planning [GM09, 204], while they describe QGs as "[...] review meetings with a “go” or “no-go” decision being made before commencement of the next phase" [GM09, 202].

Overall, we propose QGs in the context of MQG4AI that guide continuous KM through structured IM to contributing stakeholders in form of a generic and customizable lifecycle planning blueprint that addresses AI's evolutionary and dynamic character, as well as use case-specificity of concrete design decisions. The extracted information can function as foundation for meetings, as well, aiming to promote relevant knowledge to stakeholders. Therefore, on a general level, MQG4AI is envisioned as a flexible information foundation that enables further analysis based on the extracted, linked, and documented data, possibly comprising the foundation for an elaborated scoring system that contributes valuable information to AI QM.

QG Information Evaluation The root node *QG4Application* summarizes project-wide information on the AI lifecycle, comprising horizontal and vertical dependencies on all QGs combined. It represents all processes and design decisions that result in the specific intelligent system at a given point in time, aiming for flexible system evaluation scenarios. The *QG Evaluation* mechanism propagates to lower-levels and is executable for any QG scope, depending on the selected level for evaluation, following designed and identified interdependencies. For instance, some analysis may focus on the development stage only, or are tailored to the model explanation stage in particular, depending on the evaluation's objective. The resulting bi-directional graph structure from any level comprises vertical and horizontal information connections in form of collection- and leaf-QGs. The latter offering the most fine-grained scope. For instance, the extracted vertical and horizontal information flow can be analyzed with a *process-structure-matrix* (PSM) and an *information-effect-matrix* (IEM), as depicted in Figures 4.1 and 4.2, which are explained in detail in [HS09, 209].

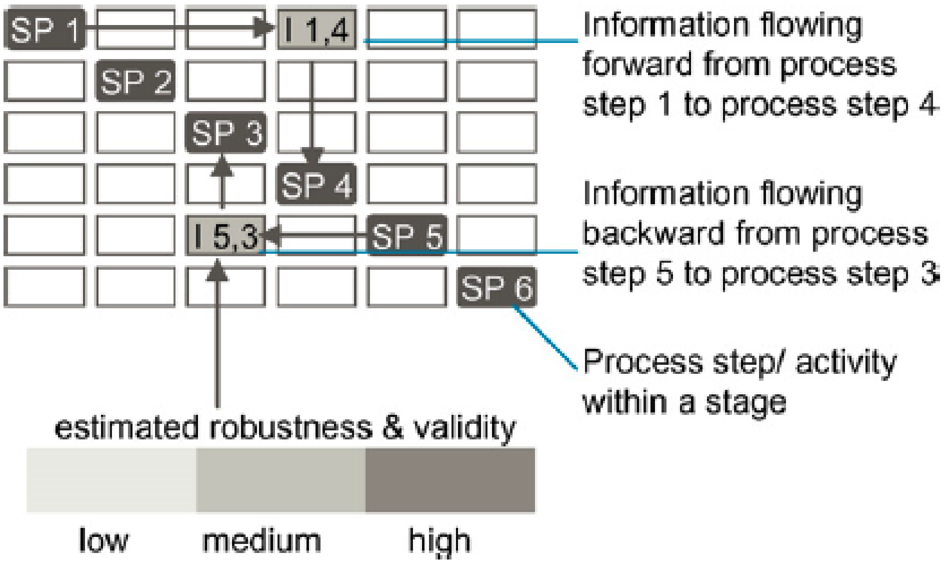

SP 1
I 1,4
SP 2
SP 3
SP 4
I 5,3
SP 5
SP 6
Information flowing forward from process step 1 to process step 4
Information flowing backward from process step 5 to process step 3
Process step/ activity within a stage
estimated robustness & validity
low
medium
high

Figure 4.1.: The *process-structure-matrix* (PSM), as depicted in [HS09, 210] to compare and evaluate information flows between processes. For instance, this concept is applicable to QGs that construct a lifecycle graph and interdependencies regarding their *input (backward information) and output information (forward information)*.

The PSM enables consistency checks for planned and documented information flows, which could serve as a tool to evaluate and compare *input (backward information) and output information (forward information)* across QGs. The information flow's quality is considered regarding parameters like the likelihood of changes, misinterpretation, and failures. Robustness of the information flow is assessed by examining factors such as the frequency, predictability, and impact of changes, as well as their propagation to dependent activities. [HS09, 210]

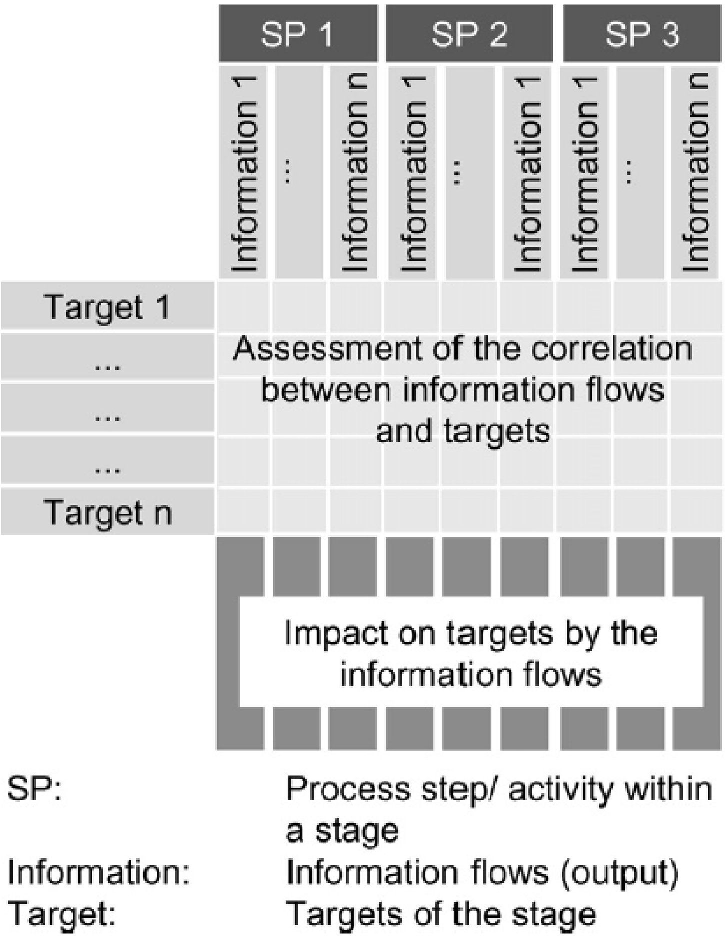

SP 1
SP 2
SP 3
Information 1
...
Information n
Information 1
...
Information 1
Information 1
...
Information n
Target 1
...
...
...
Target n
Assessment of the correlation between information flows and targets
Impact on targets by the information flows
SP: Process step/ activity within a stage
Information: Information flows (output)
Target: Targets of the stage

Figure 4.2.: The *information-effect-matrix* (IEM), as depicted in [HS09, 210] evaluates the information flow's impact on specific targets. In addition to the quality of the, with the PSM documented input and output information flows, their impact on TAI, and related risks, for instance can be evaluated. Overall, this concept can contribute to RM, and extracts a prioritization of different information flows.

In addition, the extracted information flows are assessed with an *information-effect-matrix* (IEM) [HS09, 210], as illustrated in Figure 4.2. The IEM evaluates each information flow regarding its impact on targets within a phase [HS09, 210]. For instance, targets could comprise desirable qualities of the AI system, which can be derived based on criteria for TAI, and related risks, for instance, supporting RM. In addition, IEM enables the prioritization of different information flows. Overall, this analysis of information interdependencies along the AI lifecycle in relation to the designed concept, forms the foundation for possible evaluation scenarios through IM, highlighting the calculation of different indexes, and scores that are aligned with a specific target and thus customized to specific strands of information related to (a compilation of) individual QGs.

QG Scoring System With respect to MQG4A, interpreting the thus gathered and organized AI lifecycle information is applied for customizable AI system evaluation scenarios in form of an elaborated and multi-functional *QG Scoring System* that is envisioned to consist of multiple indexes. For instance, a *Multi-Stakeholder-Fairness-Score* could be derived based on a vertical and horizontal analysis of stakeholder inclusion across all QGs within a predefined scope. In addition, a *Compliance Index*, possibly the main index for compliance assessment of the intelligent system could evaluate other scores and indexes within a single number, while simultaneously providing transparency regarding its underlying information through the MQG4AI lifecycle blueprint, and its versions with respect to MQG4A. In addition, the degree to which information within QGs is present, i.e. gate-keeping, could contribute to its calculation. For MQG4DK scenarios, the evaluation of information on RAI lifecycle concepts could serve to learn more about AI as a technology. For instance through analyzing approaches to evaluation in general and comparing them across structural similarities of multiple use cases and for different AI techniques. As a potential benchmark for AI quality criteria to compare against, an AI quality model [KV22] could be referenced, providing a systematized approach, and, based on how many criteria are addressed within the conceptual evaluation, for instance, conclusions on overall quality can be made. In addition, these criteria are equally closely related with RM and TAI.

In this section, we integrate the concept *Quality Gates* into the context of AI from a conceptual viewpoint. They comprise the foundational building block for lifecycle-adapted RAI information organization and processing in the context of the MQG4AI lifecycle planning blueprint. The next sections outline *collection-QGs* that construct the generic AI lifecycle, which is at the heart of MQG4AI, before defining the *leaf-QG* information processing structure in more detail.

4.2. High-level *Collection-QGs* – Generic AI Lifecycle Design

Building on the lifecycle flow defined by ISO/IEC FDIS 5338 on *AI system life cycle processes* [ISO23c, 36], as introduced in section 3.1.3, and drawing inspiration from the *CADAC* AI lifecycle, "[...] a comprehensive approach that addresses and

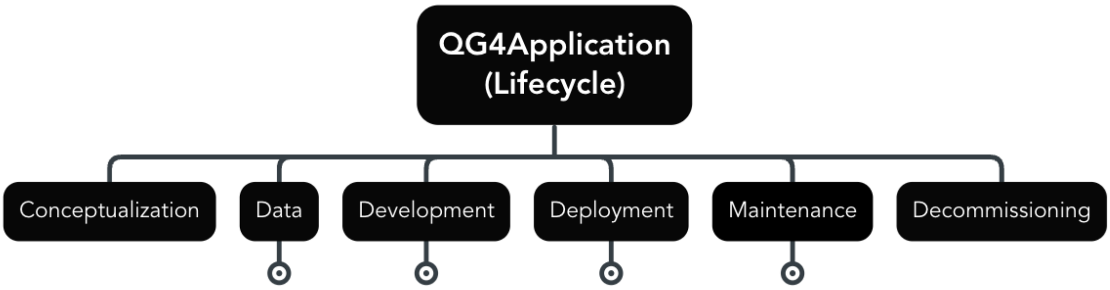


Figure 4.3.: We identify six generic AI lifecycle processes from an application view-point focusing on the organization of concepts for implementation, each representing a high-level *collection-QG* that consist of multiple sub-levels of *collection-QGs* and (compilations of) *leaf-QGs*.

accounts for all challenges from conception to production of AI" [DA22, 1], we propose a foundational high-level lifecycle structure, with the objective to capture design concepts. As shown in Figure 4.3, we define six generalizable phases (*Conceptualization*, *Data*, *Development*, *Deployment*, *Maintenance*, and *Decommissioning*) to be executed iteratively. They each comprise high-level *collection-QGs* that consist of multiple sub-levels of *collection-QGs* and (compilations of) *leaf-QGs*, aiming for generalizability of higher (sub-)levels. The proposed process structure is designed to be generalizable across diverse *AI lifecycle* design choices, starting after the *inception* phase, which involves defining contextual information such as business analysis or system requirements [ISO23c, 6]. These details are integrated into MQG4AI as *supplementary information blocks*, to facilitate information linking with the AI lifecycle. In addition, the six phases of the proposed lifecycle design accommodate varying providers for specific lifecycle stages, ensuring adaptability for a wide range of use case-specific implementations. For example, *Data* may be sourced externally for in-house *Development*, alongside externally provided *Deployment* and *Maintenance* services. Meanwhile, sub-processes such as *data preprocessing* and the system's *on-boarding strategy* in its intended real-world context must be developed internally to align with the intended purpose.

This section demonstrates the high-level *collection-QGs* that define the generic AI lifecycle in more detail, including considerations on generalizable sub-levels. They comprise the foundation to design the *lifecycle information block* of the MQG4AI lifecycle blueprint. The next section 4.3 details the proposed *leaf-QG* information processing template.

4.2.1. QG Conceptualization

During *Conceptualization*, the lifecycle is designed, and its evolutions monitored. In the context of MQG4AI, this stage is particularly relevant to MQG4A, in contrast to MQG4DK, which is based on RAI knowledge contributions on concrete implementation guidelines, and organized along the generic lifecycle. MQG4A in contrast, reflects a single AI project, and template versions are populated with QGs, as well as updated in alignment with lifecycle iterations to establish and monitor an overview of the implementation, including related contextual information. This setup aims to adhere to the adoption of "[...] flexible design, deployment and operation approaches" [ISO22a, 32], guided by RAI design knowledge based on the living lifecycle blueprint MQG4DK that forms the unifying counterpart to MQG4A. The abstract workflow to kickstart MQG4AI application during individual projects is outlined as follows, which we illustrate for a concrete application scenario in chapter 8:

1. *MQG4Application-v0* pulls a blueprint from *MQG4DesignKnowledge*, filled with information on lifecycle design that is relevant to the use case, which is configured accordingly based on *QG Tags* and an intelligent search.

2. *MQG4A-v1 to n* serve to gradually populate the individual lifecycle template with use case-specific information and connected lifecycle QGs. A project-specific branching-structure is identified, including access definition for contributing stakeholders. This may include one main concept (or, master) branch that mirrors the "final" model, and includes shared information among stakeholders. Additional, separate branches can be created by individual stakeholders to develop design concepts for separately conducted analysis, e.g. for test versions, analogously to Git-branching[4].

Overall, MQG4A reflects the AI system concept through organizing design decisions in alignment with concrete experiments and data, resulting in comprehensive lifecycle planning. Further, the blueprint is envisioned as a tool that supports verification and validation procedures in addition to contextual information linking, as well as lifecycle evolutions.

4.2.1.1. Lifecycle Phases

The demonstrated high-level structure aligns globally with the OECD's lifecycle framework [Org23, 16], which reflects ISO/IEC 22989 *Artificial intelligence concepts and terminology* [ISO22a, 36], that serves as the foundation for ISO/IEC 42001 *Artificial Intelligence Management Systems* [ISO23b, 31], and equally corresponds to ISO/IEC FDIS 5338 on *AI system life cycle processes* [ISO23c, 5]. However, the MQG4AI structure highlights lifecycle planning functionalities of design concepts, and this perspective results in some diversion, as depicted in figures 4.4 and 4.5.

[4] `https://git-scm.com/book/en/v2/Git-Branching-Branches-in-a-Nutshell`

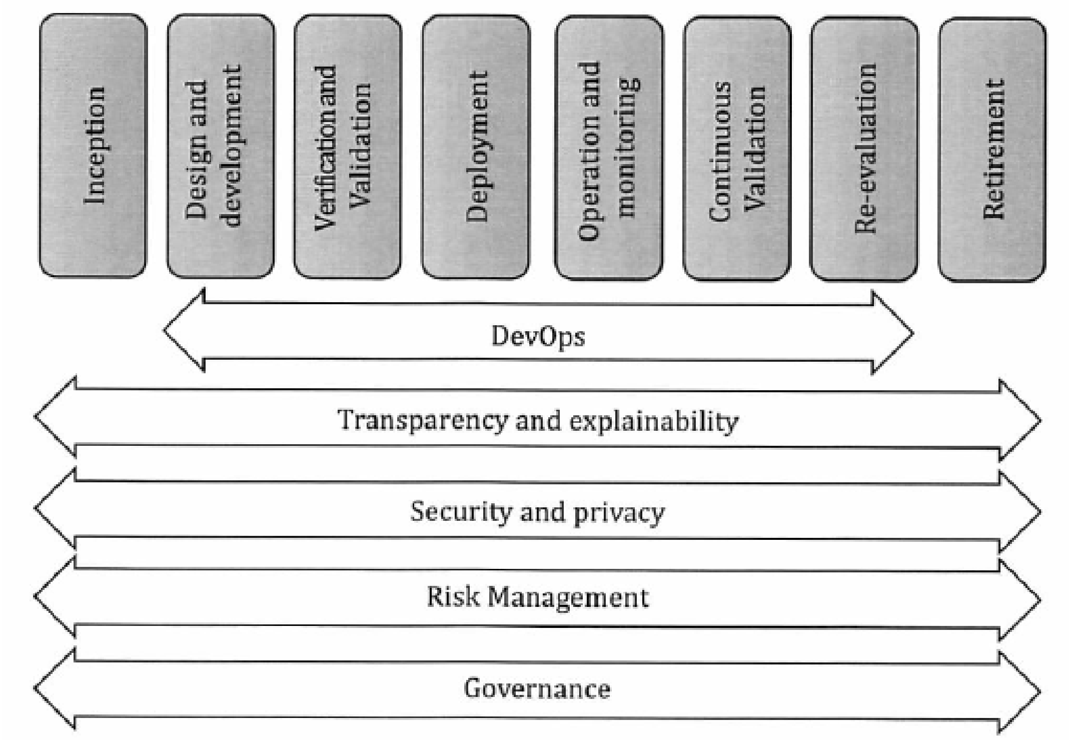


Figure 4.4.: The high-level lifecycle processes, as depicted in ISO 5338 [ISO23c, 5]. They appended *Continuous Validation*, since this stage "[...] is also applicable in situations without continuous learning, for example, to detect data drift, concept drift or to detect any technical malfunctions" [ISO23c, 5], in contrast to ISO 22989 [ISO22a, 36]. Additionally, necessary management processes and considerations throughout the lifecycle are outlined.

Specifically, we aim to answer the questions *Where is the model, and what does it need?*, providing a generic lifecycle framework designed to capture concepts for practical implementation. Regarding the depicted processes that span the lifecycle, MQG4AI's four fundamental principles highlight RM, and processes for risk mitigation (*stakeholder inclusion*, *domain embedding*, *interdependency analysis*). Within the proposed RM information block, *Transparency and Explainability*, as well as *Security and Privacy* are included with criteria for TAI. Further, the *Development* stage includes an *Explanation* phase for information monitoring of related design decisions, and *DevOps*, or *Machine Learning Operations (MLOps)* [QL24], related considerations are assigned to the *Maintenance* stage. Finally, the MQG4AI lifecycle blueprint is envisioned to support *Governance* as a contribution to AI QM through IM and KM. The lifecycle phases relate as follows:

1. The *Inception* phase, which involves defining contextual details such as business analysis or system requirements [ISO23c, 6] is embedded within *Conceptualization*. Once the project's feasibility is confirmed, application-

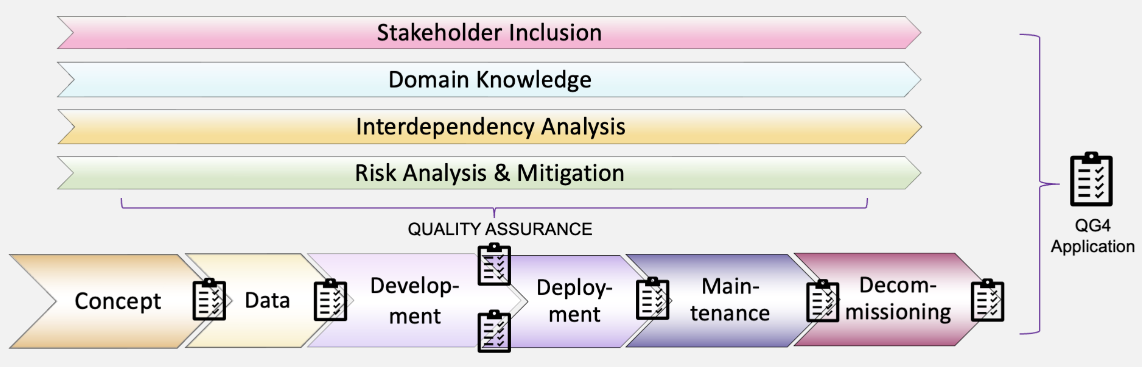


Figure 4.5.: Identified generic AI lifecycle stages, mirrored by QGs that comprise the foundation for designing the MQG4AI lifecycle blueprint, answering the questions *Where is the model and what does it need?*, and focusing on organizing implemented concepts. Their design follows four fundamental design principles, as introduced in section 3.3.2.

specific information populates MQG4AI's customizable contextual information blocks, facilitating information linking and high-quality implementation.

2. We distinguish the *Design* phase from *Development* to emphasize the continuous lifecycle process of *Conceptualization*. During *Conceptualization*, planning and ongoing updates are carried out using the MQG4AI lifecycle blueprint. These updates focus on the model and align with insights gained from other stages and contextual information.

3. Next, we explicitly emphasize the *Data* stage, which encompasses design decisions and information critical for guiding further lifecycle design choices. At this stage, the model may not yet exist or could be undergoing updates, with the quality and transformation of data continuously shaping the model's overall quality.

4. During *Development*, the model is constructed, evaluated, and refined. Unlike ISO 22989, who highlight *Verification and Validation* as standalone, quality verification and validation is implemented through a combination of related MQG4A template versions, derived from MQG4DK. Both processes support QM [ISO16] [ISO23b], and require "[...] implementing a quality management process" [ISO23c, 29]. This involves testing whether the model performs as expected [ISO22a, 39], which is typically conducted "[...] by using verification datasets [...]" [ISO23c, 27], while code review is equally relevant, if accessible [ISO23c, 27]. *Verification* aims to "[...] provide objective evidence that a system or system element fulfills its specified requirements and characteristics" [ISO23c, 26], and *validation* requires "[...] to provide objective evidence that the system, when in use, [...] achieves its in-

tended use in its intended operational environment" [ISO23c, 28]. MQG4A-template versions reflect conceptual combinations, including concrete results aligned with previously established metrics during *Development*, as well as related contextual, and other relevant information. For instance, qualities like model robustness can be assessed using adversarial examples that incorporate changes from the data stage during preprocessing and influence performance evaluation. Following the introduced branching logic, this information is organized as an MQG4A template branch for testing, and possibly some findings are integrated into the "final" model concept, the main branch, which is released into real-world deployment as a contribution to the monitoring strategy. Overall, information on MQG4A-branches is included within the *Conceptualization* stage.

5. The *Deployment* and *Maintenance* stages focus on integrating and monitoring the model in real world scenarios. These stages are closely connected to *Development* and *Data* while introducing additional layers of information tailored to the target environment. These layers include real-world data distributions and integration with existing pipelines, emphasizing the importance of MLOps that cover the complete lifecycle, as depicted in Figure 4.6. *Continuous Validation*[5] and *Re-evaluation* of the model's outputs in real-world conditions, as outlined in ISO 22989 [ISO22a, 40], play a critical role in shaping the *Maintenance* stage. Possibly, this process initiates a new lifecycle planning cycle, leveraging previously acquired knowledge that can be integrated into the updated MQG4A version. Overall, these processes are embedded within the identified conceptual interdependencies between the *Development* and *Data* stages.

6. Finally, *Decommissioning* focuses on considerations for withdrawing the model from real-world applications, potentially reversing steps from earlier stages. ISO 22989 incorporates model replacement with an updated version as part of the *Retirement* stage [ISO22a, 40], while we include this process with *Maintenance*.

In summary, this cyclic approach intends to facilitate the supervision of conceptual information on AI projects within a comprehensive AI system lifecycle, systematically organized across multiple iterations during *Conceptualization*. In addition to the described *roadmap* [GM09] for QG positioning, monitoring of different MQG4A template versions contributes to reliable design choices along the AI lifecycle that address AI's intricacies, enable on-going organization of complex testing strategies, including enhanced traceability of stochastic outputs, through the execution of multiple iterations [ISO22a, 36], among other IM-based opportunities.

[5] This stage "[...] is also applicable in situations without continuous learning, for example, to detect data drift, concept drift, or any technical malfunctions" [ISO23c, 5].

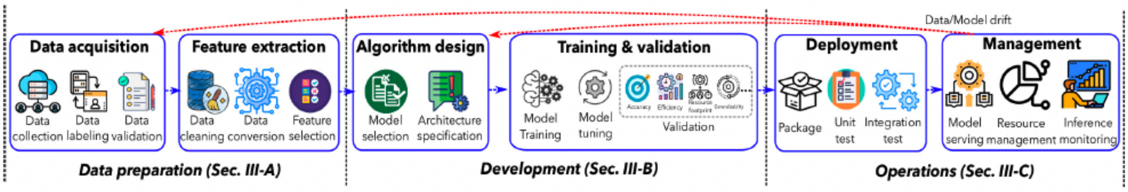


Figure 4.6.: As an additional perspective, the lifecycle from an *Machine Learning Operations (MLOps)* viewpoint, as depicted in [QL24, 3] closely aligns with our proposed conceptual structure.

4.2.1.2. Design Decision-Making

Generally, the scope of *Design Decisions* within MQG4AI is intentionally defined as flexible to enable customization based on specific project contexts. For instance, when using automated, state-of-the-art pipelines during e.g. model configuration, several sub-decisions, such as those related to architecture selection or pre-processing, may be handled externally. As a result, fewer corresponding leaf-QGs are instantiated, since detailed design steps are effectively outsourced. This flexibility supports a scalable application of the blueprint across diverse levels of system complexity and automation.

During individual projects, different versions of the MQG4A-template incorporate concrete results, reflecting the evolution of individual lifecycles with respect to the design decision-making process. For example, MQG4A versions can be categorized into *pre-selection*, *intra-selection*, and *post-selection* stages for lifecycle planning. This approach is based on our empirically derived design decision-making method aimed at reliable performance evaluation metrics for a fictional use case in emergency medicine (EM) [EM24], which can be tested and refined to guide relevant design choices throughout model development. Through a retrospective analysis, accordingly structured MQG4A template versions are simulated in chapter 8, based on the medical use case introduced in section 5.2.2. Figure 4.7 highlights the intricacies of decision-making. Focusing on interdependencies that influence model optimization, the iterative design decision-making process showcases generated results and their updates for interrelated model configurations, explanations, and preprocessing steps, all grounded in data quality and a robust performance evaluation strategy. Finally, aiming to advance shared AI design, MQG4AI comprises a module to document knowledge on design decision-making as part of the *model development* stage.[6] In addition, information on public data sets that contribute to public bench-marking of the proposed approach is included. This setup is intended to facilitate scientific publications to further enhance global RAI knowledge (MQG4DK), as depicted in Figure 3.14.

[6] `https://github.com/miriamelia/MQG4AI/tree/main/MQG4DesignKnowledge/2_Lifecycle/2_Development/0_DesignDecisionMaking`

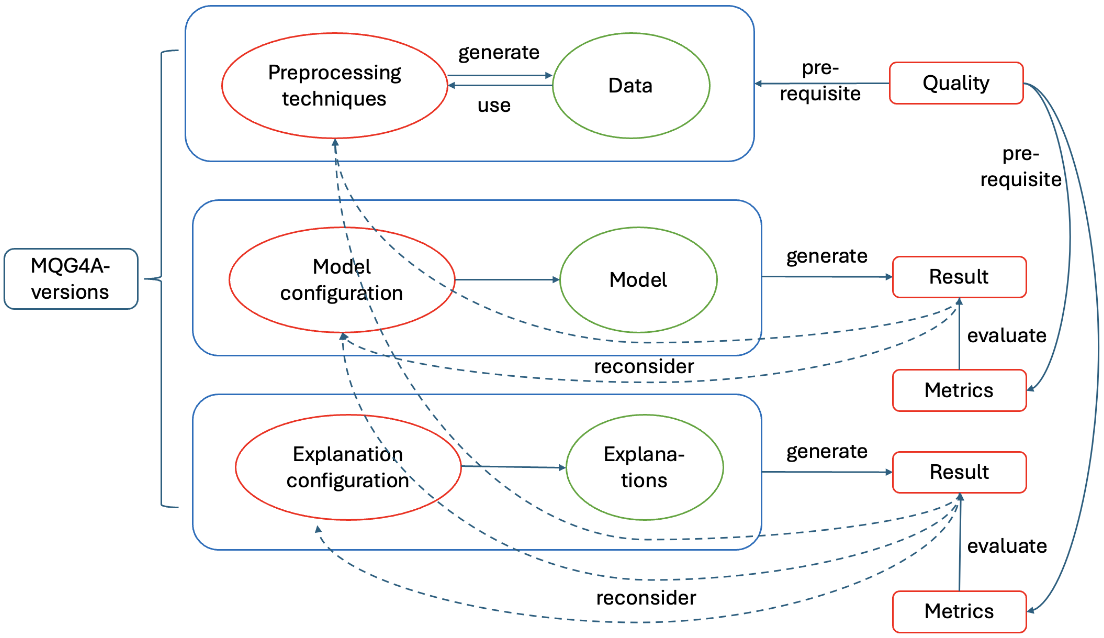


Figure 4.7.: The iterative process of configuring the model, including hyperparameters, and the explanation method is depicted. Situated during the optimization step in the development stage, we focus on outputs and interdependencies. These design decisions are influenced by preprocessed data and are evaluated against predefined metrics that need to be tailored to the specific context and structural setting. These design choices span different stages and must be carefully planned. A combination of MQG4A versions is proposed to structure the iterative design decision-making process as part of a continuous lifecycle conceptualization. For example, MQG4A versions can be organized into *pre-, intra-, and post-selection steps*. Based on the necessary information gathered prior to making the design decision, such as evaluation metrics, evolving results are compared with adjustments to a baseline approach until reasonable updates are achieved. This interaction scenario is illustrated in chapter 8 for a concrete medical use case.

The following sections outline proposed high-level collection-QGs for the generic lifecycle stages in more detail. Based on our experience, we concentrate on abstracting generalizable process steps for the *Development* stage and the related *Data Utilization* as starting point to kick-start MQG4AI design.

4.2.2. QG Data

The generic data lifecycle stage comprises three main sub-processes, as depicted in Figure 4.8: data *Acquisition*, *Utilization*, and *Maintenance*, while the latter are strongly related to model *Development* and *Maintenance* (possibly including *Decommissioning*), respectively. The structure is derived from the *IEEE Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence* [IEE22], and organizes the proposed data lifecycle that consists of 11 steps in alignment with the *CADAC lifecycle* [DA22], and the *AI data engineering process*, as illustrated in ISO 5338 [ISO23c, 22]. Overall, data plays a crucial role when training and evaluating an AI model, and it is important to "[...] understand its source, how it was processed, its owner and its rationale [...]" [ISO23c, 24]. Additionally, the EU AI Act requires documented information on the data, and a comprehensive data governance, as defined in Article 10 [Fut24].

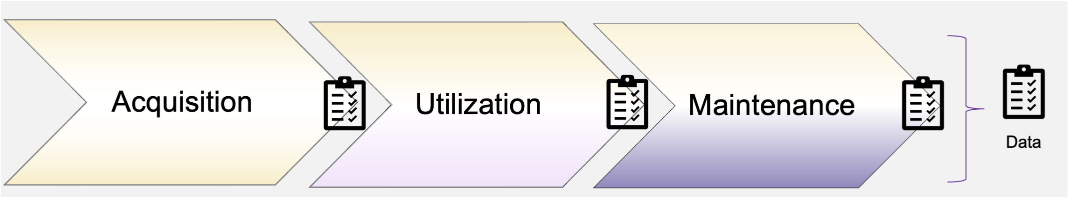


Figure 4.8.: Highest sub-level of *Collection-QG Data*. Data *Utilization*, and *Maintenance* strongly correlate with model *Development* and *Maintenance* (possibly including *Decommissioning*).

The following sections introduce more fine-grained considerations on related sub-design choices for data *Acquisition*, *Utilization*, and *Maintenance*, while our practical experience mainly relates to the data *Utilization* phase.

4.2.2.1. QG Acquisition

Data acquisition is organized into multiple sub-levels that are carried out by the organization responsible for the data set. From the model development viewpoint, external data acquisition might be necessary, and in that case, "[...] it is recommended to develop a long-term data acquisition strategy based on trustworthy relationships with all stakeholders by being transparent on what data are collected and how they are used" [DA22, 6]. Overall, the data needs to be aligned with the intended use, and the data should reflect the dynamics of the real-world setting [ISO23c, 23]. Further, it "[...] must be cross-examined and validated for potential security risks (linking back to the preliminary risk assessment), and ethical and legal conformity" [DA22, 5]. The following list comprises necessary process steps, including the exploration of quality guidelines. Their execution may happen iteratively, and in parallel, and is depicted in Figure 4.9.

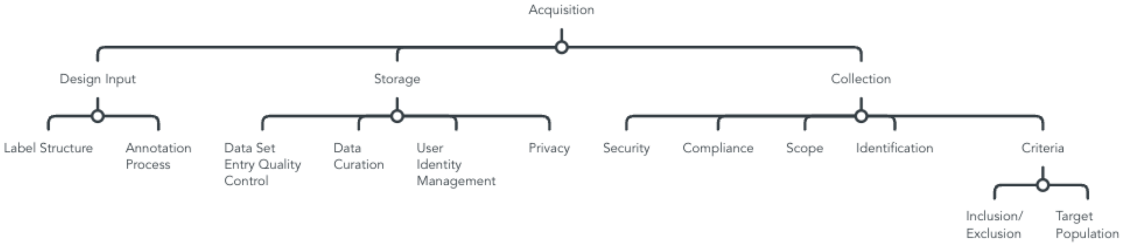


Figure 4.9.: Exemplary overview of data acquisition processes and design decisions.

QG Design Input A robust dataset begins with a well-defined design aligned with the intended use, ensuring (clinical) representativeness, clear annotation processes, and appropriate data formats [IEE22, 17], which may include domain-embedded data labeling, as a "special form of data acquisition" [ISO23c, 23]. Key steps comprise documenting the annotation task, its outputs, rules, and verification of results, as well as detailing the annotation process, including decision-making criteria and annotator qualifications [IEE22, 20]. A controlled annotation process with competent individuals ensures quality and consistency. Finally, data exploration supports domain understanding, ensuring the dataset meets application-specific needs. [ISO23c, 23]

QG Data Storage Effective data storage is essential for reliable dataset management. It involves keeping records of permission control, user activity, data utility, and data distribution, ensuring privacy, user identity management, and reliable storage systems with backup and repair capabilities [IEE22, 20]. A unified data repository, such as a data warehouse, data lake, or data lakehouse, is recommended over data silos to centralize data access, ownership, stewardship, metadata, ethics, governance, and regulatory compliance [DA22, 5]. This enables efficient data curation, "[...] i.e., the process to compile the data, actively manage and add value to the data, and realize the transformation from data collection to data processing and use" [IEE22, 19], including processes like integration, aggregation, and packaging, while maintaining transparency and monitoring data statistics [IEE22, 19]. Quality control actions by dataset administrators are crucial before data entry to ensure high-quality data [IEE22, 20], supporting continuous reliable AI model creation.

QG Data Collection Data collection requires well-defined standard operating procedures to ensure data uniqueness, traceability, and security. These include criteria for data collection (e.g. inclusion/exclusion conditions, target population) and secure handling practices. [IEE22, 18] For a reliable data distribution, operating procedures must enforce fine-grained authorization, consistency monitoring, and secure import/export to maintain security and integrity [IEE22, 21]. Protecting sensitive data is paramount, as AI systems face increased attack surfaces with augmenting data sources, particularly during data storage and transfer [ISO23c, 24]. Employing privacy-preserving techniques [ISO23c, 24] and addressing data privacy, confidentiality, and integrity are critical. These aspects, often governed by regulations like the GDPR, help mitigate risks from poisoning and backdoor attacks [DA22, 5].

4.2.2.2. QG Utilization

The data utilization phase strongly correlates with the AI model, as it cleans, merges, and prepares the data with the aim to extract input features [ISO23c, 24], as well as to ensure data quality. Analyzing data quality is an on-going process [ISO23c, 23], and aiming for data that is "fit for purpose" [DA22, 6] includes continuous (automated) quality checks and verification procedures [ISO23c, 23].

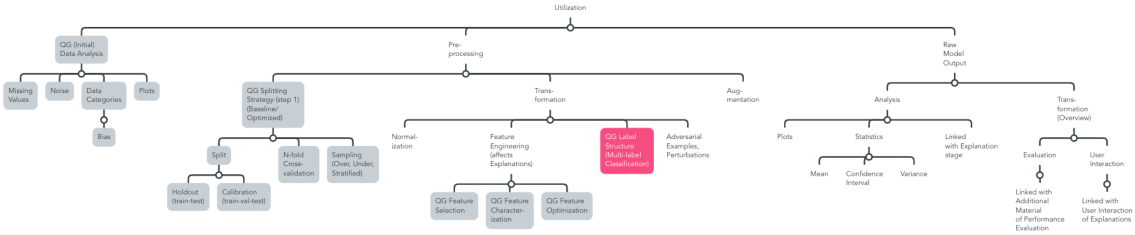

Utilization
QG (Initial) Data Analysis
Missing Values
Noise
Data Categories
Plots
Bias
Pre-processing
QG Splitting Strategy (step 1) (Baseline/ Optimized)
Split
N-fold Cross-validation
Sampling (Over, Under, Stratified)
Holdout (train-test)
Calibration (train-val-test)
Trans-formation
Normal-ization
Feature Engineering (affects Explanations)
QG Feature Selection
QG Feature Character-ization
QG Feature Optimization
QG Label Structure (Multi-label Classification)
Adversarial Examples, Perturbations
Aug-mentation
Raw Model Output
Analysis
Plots
Statistics
Mean
Confidence Interval
Variance
Linked with Explanation stage
Trans-formation (Overview)
Evaluation
Linked with Additional Material of Performance Evaluation
User Interaction
Linked with User Interaction of Explanations

Figure 4.10.: Proposed sub-level QGs for the data utilization phase. Leaf-QGs are marked gray, and the pink contribution belongs to a leaf-QG compilation for reliable performance evaluation metrics, as illustrated in chapter 7 for multi-label ECG classification.

In addition, data preparation is typically exploratory and ad hoc, making it challenging to repeat. To address this, aiming for reusable and automated preparation is beneficial. Due to its complexity, (automated) testing is crucial to ensure reliability. [ISO23c, 24] Capturing implemented concepts with the MQG4AI lifecycle blueprint supports this exploratory and iterative process, and overall, these design choices, as illustrated in Figure 4.10, impact the preparation of high-quality data that ensures accurate, efficient, and reliable AI model behavior.

QG (Initial) Data Analysis Initial data analysis focuses on cleaning and visualizing data to ensure it is suitable for its intended use [IEE22, 18]. This process involves identifying and correcting errors such as missing values, noise, outliers, duplicates, and inconsistencies, while adhering to inclusion and exclusion criteria that align with the dataset's purpose and data collection requirements [IEE22, 19]. Key steps include exploring relationships between data points and attributes to uncover patterns or anomalies [DA22, 5] and conducting data quality analysis to ensure bias remains within acceptable limits [ISO23c, 23]. Additionally, addressing risks such as data poisoning, which can lead to undesired changes in model behavior, is critical [ISO23c, 23]. Filtering plays a significant role during this phase, removing data that is irrelevant, excessive, or harmful. Examples include the exclusion of biased or discriminatory data, de-identifying personal information to comply with privacy regulations, and protecting sensitive data from unauthorized access. Filtering also ensures the dataset is free from elements that could violate legal or ethical standards. [ISO23c, 24] The phase's goal is to produce a dataset that is clean, compliant, and robust, providing a solid foundation for subsequent model development and evaluation.

QG Preprocessing The pre-processing stage is essential in ensuring that all acquired data for building AI models accurately functions as input for algorithms, with minimal loss of accuracy, informational value, and data quality [DA22, 6]. This stage includes several key components. One important consideration is *splitting the data* for model development and evaluation [DA22, 6]. Utilized test data needs to come from a different source than training data to mitigate generalization issues and enhances model reliability [ISO23c, 23] [MF22, 3]. In addition, "it is expected to be similar to production data [...]" [ISO22a, 21]. Further, *data splitting* needs to be executed before *data transformations* to avoid data leakage and the independence assumption [MF22, 8]. Also, splitting strategies such as n-fold cross validation, and sampling techniques may influence one another, and their application needs to be correctly evaluated [BL15]. *Data transformation* involves adjusting the values and characteristics of data to meet its intended use. This may include normalization, data enhancement, and feature engineering [IEE22, 19]. For instance, this includes the conversion of non-numerical data types into numerical representations such as label encoding or one-hot encoding [DA22, 6]. Or, *normalization*, for example, standardizes units of measurement or date formats across all variables to ensure consistency [DA22, 6]. *Feature engineering* focuses on selecting, optimizing, and characterizing features for AI models. It leverages domain knowledge and data analysis to minimize risks of errors by using fewer, more effective input features. [ISO23c, 24] This process is strongly related to model explanations that explain the selected features of individual predictions, for instance. Finally, *data augmentation* can enhance training data through techniques like class imbalance management, feature engineering, and representation [DA22, 7]. It comprises the selection of an augmentation technique, which may be based on a DNN, such as a generative adversarial network (GAN), which may include additional training data, as well as a verification procedure of the applied methods [IEE22, 22]. Unlike transformation, which focuses on variables, augmentation involves the model being iteratively re-run, and necessitates an (on-going) evaluation of the effectiveness of these measures [DA22, 7].

(Raw) Model Output Finally, we append an overview of implemented concepts on raw model output to the data utilization stage. In addition to analyzing the raw model output, which may include the creation of plots, and the calculation of statistical values, such as mean, variance, or confidence intervals, adding an additional perspective on the opaque model's performance, we are convinced that providing a summary of raw output transformations is beneficial for lifecycle monitoring. These analysis may provide relevant information to the explanation stage, where raw model outputs are formalized and visualized. Further, transforming the raw output for performance evaluation, e.g. in form of a threshold-based confusion matrix for classification metrics, or providing information on targeted output transformations to the user, which needs to be aligned with the envisioned user interaction of explanations, is summarized within the (raw) model output stage. Possibly, with more MQG4AI testing, this information module may be assigned to the model development stage.

4.2.2.3. QG Maintenance

Finally, the data changes during project execution in the real world, which necessitates maintenance processes. For instance, the data is subject to technical and ethics reviews post-deployment, as well as a crucial consideration when building pipelines for operationalization, which includes continuous monitoring of the data's evolutionary character that impacts model performance [DA22, 9].

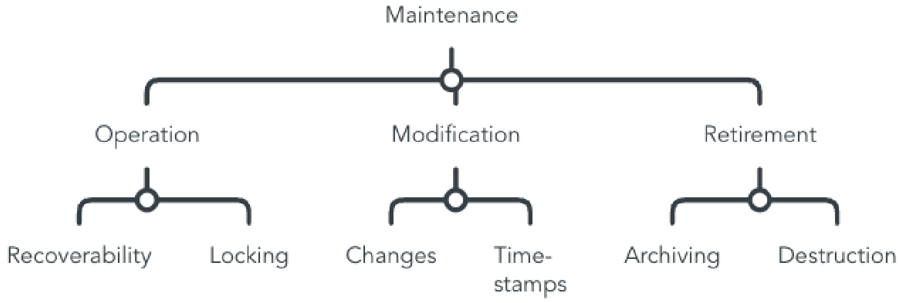


Figure 4.11.: Sub-level QGs for the data maintenance phase are illustrated.

Therefore, data maintenance is closely related with model maintenance [ISO23c, 28] [ISO23c, 32] [ISO23c, 33], and these practices collectively contribute to maintaining a secure and reliable data environment throughout the deployment and operational stages. An exemplarily structured view on sub-processes is depicted in Figure 4.11.

QG Operation Data operationalization involves crucial aspects such as recoverability and locking to safeguard data integrity and security. Recoverability refers to the ability to backup and recover data sets effectively, ensuring data is preserved in the event of failures or accidental loss [IEE22, 21]. Locking of data sets plays a vital role in preventing unauthorized operations, such as modification, deletion, or access, ensuring that data remains secure and tamper-proof [IEE22, 21]. Additionally, privacy protection and cybersecurity measures, such as user authorization and encryption methods, are essential to uphold data security and safeguard sensitive information [IEE22, 22].

QG Modification Managing data modification involves establishing standard operating procedures to ensure data updates are handled effectively. Modifications should be documented meticulously for auditing purposes, including pre- and post-modification conditions, timestamps, executors, inspectors, interfaces, reasons for changes, and the results of modifications. Additionally, whenever a data set is updated, the responsible organization should notify users through appropriate channels, such as email or service change subscriptions, while maintaining adherence to established quality characteristics. [IEE22, 22] This process ensures transparency and accountability in data management and enhances the overall integrity of data throughout its lifecycle.

QG Retirement Finally, data retirement is executed when a user makes a request, when a contract ends, or when a data authorization agreement is ended. The process involves preparing for data archiving, as well as the secure destruction of data and storage mediums to maintain privacy and data security. [IEE22, 23] Whether executed in conjunction with model decommissioning or independently, careful management of data retirement is essential to uphold data integrity and regulatory compliance.

4.2.3. QG Development

The model development lifecycle stage is organized into four sub-process, namely model *Configuration*, *Evaluation*, *Optimization*, and *Explanation* that cover implementation concepts. All processes are executed in an iterative manner, and they are highlighted in Figure 4.12. This stage comprises the main focus of our contribution.

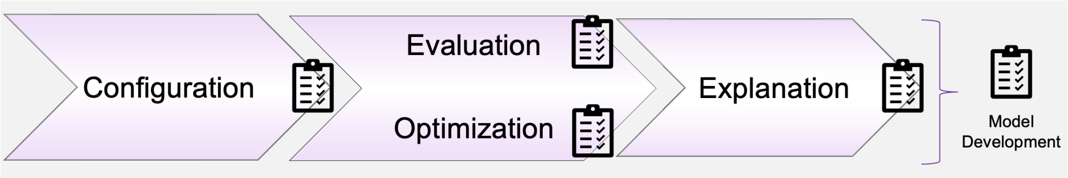


Figure 4.12.: Sub-processes that contribute to the *collection-QG Development*. Note the continuous interplay between *model optimization* and comparative output *evaluation* for design decision-making, contrasting a baseline configuration with implemented concepts for testing towards model optimization. This is relevant for MQG4A scenarios, while MQG4DK contributions are organized within the proposed generic structure.

For instance, concepts for optimization address the initial configuration, and propose changes, which includes building multiple models and their comparison against a benchmark [DA22, 4]. Implemented concepts regarding model explanation are clustered separately. Overall, the proposed structure is aligned with sub-processes of the *CADAC* lifecycle [DA22], and the *implementation process*, as outlined in ISO 5338 [ISO23c, 25], which results in "[a] working AI model [...]" [ISO23c, 25], including "[d]ocumentation of the modelling process [...]" [ISO23c, 25], for instance in form of MQG4A. The aim is to "[...] determine an overall approach to designing the AI system, testing it and making it ready for acceptance and deployment" [ISO22a, 38].

Generally, AI system creation extends beyond traditional software engineering, and AI introduces novel elements that require systematic management in a *backlog* [ISO23c, 26], in response to which MQG4AI offers a comprehensive method. This supports "[...] cross-disciplinary coordination, planning and evaluation" [ISO23c, 26], aiming to facilitate model training and "[...] establish an internal representation [...]" [ISO23c, 25] of the AI lifecycle. The process is iterative, combining and relating configuration, evaluation, optimization, and explainability considerations. Modern development tools, particularly Jupyter notebooks and feature-oriented algorithm libraries, as well as repositories like GitHub, have significantly enhanced the efficiency of AI model development [DA22, 7]. In addition, throughout this stage, "[...] processes and controls described in the risk treatment plan" [ISO22a, 39] to ensure proper RM should be implemented. In the context of MQG4AI, concrete design decisions are linked with the RM information block in form of the *leaf-QG* template, which is introduced in section 4.3.

4.2.3.1. QG Configuration

This section summarizes concepts on how the first model [DA22, 7], or the model that is being evaluated against optimization concepts in later iterations, is designed. This broadly includes information on the model *architecture*, *hyperparameter*, as well as the *training strategy*, as depicted in Figure 4.13.

Overall, the selection and optimization of algorithms involves experimentation to determine the most suitable technique [ISO23c, 25], which is a form of optimization, and relates to concepts for optimizing initial design decisions. Also, model selection includes considering the trade-off between interpretability and performance [ISO23c, 25], for instance. Another trade-off relates to *transfer learning*, where "[...] an existing machine learning model is used as a starting point to further train for a slightly different use case" [ISO23c, 26]. While pre-trained models are available and may save resources, reimplementation can ensure model integrity and address trustworthiness concerns related to both the model and its source data through in-house development [DA22, 7].

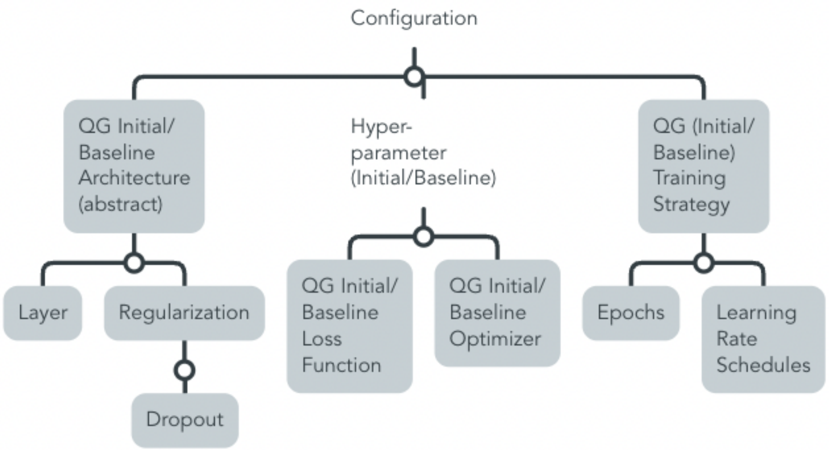

Configuration
QG Initial/ Baseline Architecture (abstract)
Hyper-parameter (Initial/Baseline)
QG (Initial/ Baseline) Training Strategy
Layer
Regularization
Dropout
QG Initial/ Baseline Loss Function
QG Initial/ Baseline Optimizer
Epochs
Learning Rate Schedules

Figure 4.13.: Relevant information for *model configuration* is broadly separated into *architecture*, *hyper-parameter*, and *training strategy*. This includes information on design decisions, such as number and type of model layers within an architecture, or the learning rate scheduling strategy during training. Proposed *leaf-QGs* are colorized gray, and their content is defined in alignment with the concrete use case in case of MQG4A, or a concrete AI technique/design decision-making compilation for MQG4DK.

4.2.3.2. QG Evaluation

Defining the evaluation strategy in alignment with the use case at hand, including data utilization steps results in developing a benchmark [DA22, 7], which functions as a directive for conducting refinements through model optimization in form of comparing multiple models with different designs.[7] A fundamental step to execute benchmarking comprises a solid performance evaluation metrics compilation, in addition to other output assessment strategies.

Performance Establishing an evaluation benchmark is essential from the start of developing the first AI model. This benchmark may be derived from human expertise, and/or from other existing knowledge surrounding the same or a similar use case, as well as standards within the relevant industry or algorithm type. Beyond evaluating the model's performance, the benchmark helps identify mismatched input attributes or incomplete data representation, guiding subsequent iterations like building additional models or refining previous data preprocessing steps. [DA22, 7] Performance evaluation metrics are compared against the previously established performance benchmark, emphasizing the need to understand what each metric represents and how it is calculated in alignment with relevant domain knowledge. Resources like Application Programming Interface

[7] For instance, the benchmarking process can be conducted collaboratively with the global community when using open-source data, as previously considered in section 4.2.1.

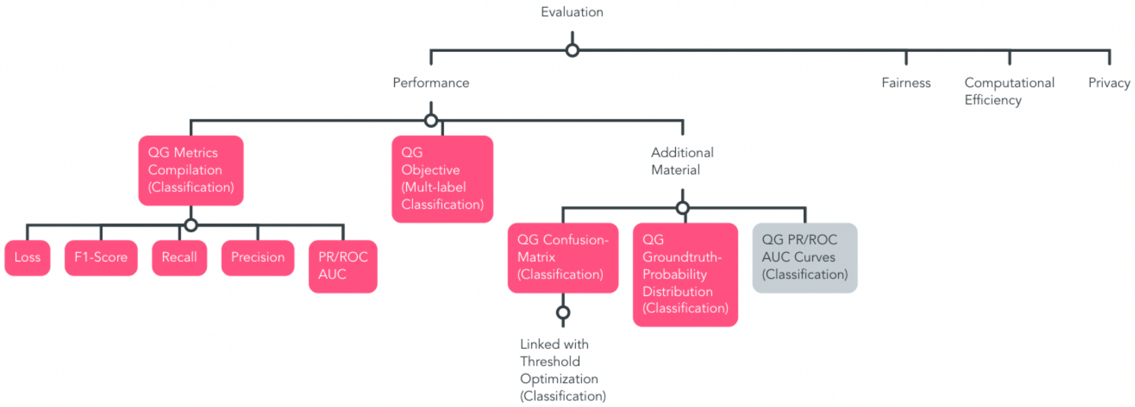

Evaluation
Performance
Fairness
Computational Efficiency
Privacy
QG Metrics Compilation (Classification)
QG Objective (Mult-label Classification)
Additional Material
Loss
F1-Score
Recall
Precision
PR/ROC AUC
QG Confusion-Matrix (Classification)
QG Groundtruth-Probability Distribution (Classification)
QG PR/ROC AUC Curves (Classification)
Linked with Threshold Optimization (Classification)

Figure 4.14.: *Performance evaluation* comprises a crucial component of model evaluation and builds the foundation for many subsequent design decisions. Therefore, it should be aligned with the use case at hand. In addition, the desired model quality in the real world includes further criteria that need to be evaluated. Proposed *leaf-QGs* are colorized gray, and their content is defined in alignment with the concrete use case in case of MQG4A, or a concrete AI technique/design decision-making compilation for MQG4DK. Pink contributions are outlined in chapter 7, corresponding to the proposed reliable classification metrics compilation, which addresses the *technical robustness*-related risk *unreliable performance evaluation metrics*.

(API) documentation and case studies provide valuable insights. Effective metrics should be accurate, robust, agnostic, scalable, and interpretable. [DA22, 8] Model verification includes checking if any "[...] relevant performance characteristic of the AI system meets specific requirements" [ISO22a, 39], emphasizing the need for reliable performance evaluation metrics tailored to the use case at hand that avoid related AI pitfalls [TJ21] [HSA22]. Common classification metrics include accuracy, precision, recall, F1 score, and root mean-squared error, among others. The generalized formula, *Outcome* = *model* + *error*, underscores that accuracy is not the sole metric to consider. [DA22, 8] For instance, "ISO/IEC TS 4213:2022 Information technology — Artificial intelligence — Assessment of machine learning classification performance"[8] [ISO22b] provides a comprehensive list of key concepts in classification performance evaluation, including statistical tests and different types of classification. Complementing this RAI design knowledge collection, in chapter 7, we outline a possible method how to approach reliable performance evaluation metrics selection for multi-label ECG classification surrounding a fictional use case situated in EM [EM24] as a contribution to MQG4DK. Overall, evaluation metrics form the foundation for subsequent design decisions, including model comparison across parameter settings for optimization, and assessing the bias-variance trade-off regarding adequate model complexity [DA22, 8].

[8] No European CEN version was found as of December 2024.

Additional Evaluation Scenarios In addition to demonstrating "intelligence" through one or more AI capabilities, an AI model must also be computationally efficient for broader deployment. This involves calculating supplementary metrics, including computational (CPU) and memory performance, time complexity, ethical implications, and convergence metrics. Additionally, metrics for assessing risks are required, assessing privacy-, cybersecurity-, trust-, robustness-, explainability-, interpretability-, usability-related impacts, as well as societal consequences, reflected by e.g. AI fairness metrics and privacy evaluation stratgies. [DA22, 8] Depending on the chosen approach, MQG4AI offers (combinations of) different approaches for evaluation. If concrete implementation concepts, e.g. in form of specific metrics, exist, MQG4A, as well as MQG4DK offer customizability for concept documentation, contributing to overall transparency. If the individual evaluation approach consists of a combination of design decisions, e.g. in combination with changes to the underlying data and monitoring of performance evaluation, such as in case of robustness evaluation through adversarial examples, different MQG4A-template versions including concrete results are monitored. With respect to MQG4DK, a combination of design decisions that belong together is appended.

4.2.3.3. QG Optimization

Model tuning involves "[e]mploying optimisation techniques to find the hyperparameters that provide the best performance, using validation data" [ISO23c, 25]. This process builds upon the performance benchmark, which helps identify gaps in the initial model. Subsequent models are then refined by incrementally increasing their complexity [DA22, 7]. A proposition to organize related design choices is depicted in Figure 4.15.

Optimization encompasses both *model selection* and *hyperparameter* configuration, iterating until the model performs adequately on the training dataset. This process is closely tied to data utilization strategies applied to validation and test data, as discussed previously. Further, expert knowledge plays a critical role in this phase, with "[...] distributing the work between experts and computer resources to experiment in parallel" [ISO23c, 26] further enhancing development efficiency. [ISO23c, 26] Concepts that optimize the model may include *post-processing* methods, e.g. tuning thresholds for performance evaluation in alignment with the data distribution. In cases where current algorithms lack the required capability, new ones may need to be created. Such efforts are undertaken only when state-of-the-art methods fail to address technical or computational challenges. These novel algorithms must be evaluated separately against benchmark datasets to validate their utility and value addition. [DA22, 7] With respect to MQG4A-scenarios, concepts for optimization are documented in an iterative manner through different versions, and multi-stakeholder template-access, enhancing optimization.

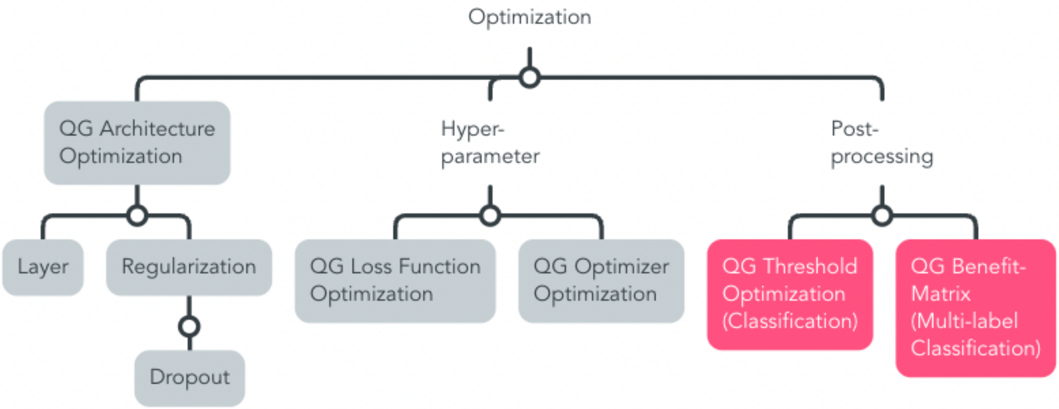

Optimization
QG Architecture Optimization
Hyper-parameter
Post-processing
Layer
Regularization
Dropout
QG Loss Function Optimization
QG Optimizer Optimization
QG Threshold Optimization (Classification)
QG Benefit-Matrix (Multi-label Classification)

Figure 4.15.: *Optimization* comprises concept adaptations of the model architecture and hyperparameters. This fine-tuning aims to keep/surpass the defined bench-mark, as identified based on the evaluation strategy. Additionally, post-processing methods after the model is trained are included. Proposed *leaf-QGs* are colorized gray, they are finalized in accordance with the concrete use case for MQG4A scenarios, which may equally comprise monitoring different template versions to assess and optimize desirable quality criteria. Concrete AI technique/design decision-making compilations are appended to MQG4DK. The pink contributions belong to the proposed reliable classification metrics compilation, and exemplify a complex MQG4DK contribution, as outlined in chapter 7.

4.2.3.4. QG Explanation

Explainability refers to an AI system's ability to present the key factors influencing its decisions in a manner comprehensible to humans. This property becomes particularly critical when AI decisions directly impact individuals (e.g. CDSS), where a lack of understanding may lead to distrust. [ISO22a, 29] Consequently, model explainability, highlighting XAI methods, has emerged as a response to ethical and regulatory demands for transparency in AI [DA22, 8]. "Explainability can also be a useful means of validating the AI system, even where the decisions do not directly affect humans" [ISO22a, 30]. It aids stakeholders by clarifying how attributes and parameters influence the system's outcomes, thereby enhancing model creation [DA22, 8]. While rule-based algorithms, such as symbolic methods or decision trees, are often considered highly explainable, their understandability can diminish as model size and complexity increase [ISO22a, 30]. DNNs are opaque in nature, necessitating the consideration of XAI methods throughout the AI lifecycle. Their intricacies, as detailed in section 2.2.2, can make it challenging to provide meaningful explanations of how decisions are made. Nevertheless, historical counterexamples demonstrate the risks of inadequate and flawed decision-making, which was uncovered thanks to XAI. [ISO22a, 30]

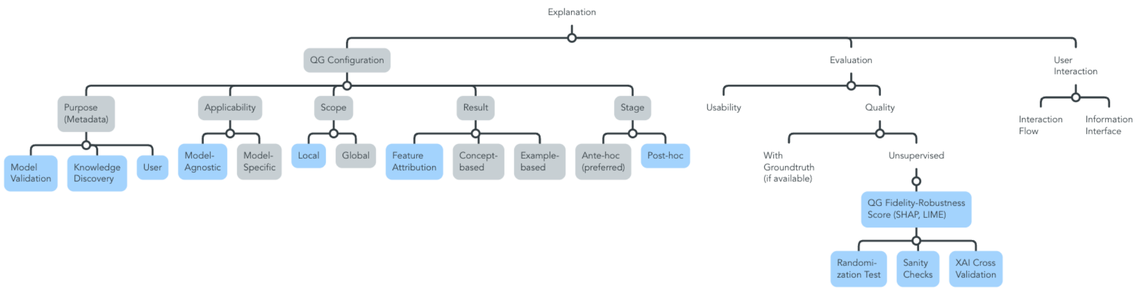


Figure 4.16.: The proposed *Explanation* stage is broadly divided into the three subprocesses *configuration*, *evaluation*, and *user interaction*. *Leaf-QGs* are depicted in gray, and they are filled with respect to concrete use cases for MQG4A scenarios, possibly including monitoring of different template versions to assess and optimize different versions for testing. Concrete AI technique/design decision-making compilations are appended to MQG4DK. The blue colorization covers relevant method configuration information, exemplifying a contribution based on a technical method for evaluating SHAP and LIME explanations [LGC24], as outlined in chapter 6.

The proposed lifecycle includes an explanation stage, which comprises method *configuration*, and *evaluation* concepts, in addition to *user interaction* steps regarding the system's application in the real world, as depicted in Figure 4.16. It is alignable with XAI concepts, as introduced in the *IEEE Guide for an Architectural Framework for Explainable Artificial Intelligence* [Art24]. The explanation stage in the context of MQG4AI, including linking with related risks is demonstrated in more detail in chapter 6, where we focus on introducing the *leaf-QG* template in the context of MQG4DK.

QG Configuration Overall, XAI methods are categorized into intrinsic (ante-hoc) and extrinsic (post-hoc) approaches based on the relevant lifecycle *stage* for their implementation. Intrinsic methods are embedded within the AI algorithm, functioning alongside the system's capabilities. [DA22, 8] However, "[m]ost AI algorithms have not been designed (or evolved) with intrinsic methods, therefore XAI is primarily a collection of extrinsic methods, such as partial dependence plots, individual conditional expectation, local interpretable model explanation, and Shapley addictive explanations (SHAP)" [DA22, 8]. Other criteria for method configuration include the method's *purpose*, its *applicability* regarding specific types of models or globally, the *scope* of the explanation, as well as the characteristics of the generated *results*.

QG Evaluation Assessing explanations in AI systems requires a focus on both *usability* and *quality*, as explanation techniques that are effective in one context may fail to meet the specific demands of another [ZJ21b]. However, evaluating explanations is a complex task, hindered by domain-specific constraints, the diverse range of properties to consider, and the challenge of incorporating case-specific metrics. Refer to section 2.2.2 for a comprehensive overview. This complexity is further compounded by the absence of generalizable methods for automating the evaluation process [AMJ18]. A possible approach to this challenge is highlighted in [LGC24], which identifies robustness and fidelity as two key, interrelated properties of explanations. The study proposes computational methods for systematically evaluating SHAP and LIME explanation techniques along these dimensions, providing a structured framework for assessing their applicability, which we append as a *leaf-QG* to MQG4DK in chapter 6 as a RAI knowledge contribution addressing the *Transparency*-related risk *unfaithful explanations*.

QG User Interaction Finally, the effective presentation and clear communication of explanations to relevant stakeholders are critical processes for elucidating the decision-making processes behind specific AI outputs. This stage emphasizes the design and delivery of explanations, leveraging user studies and usability testing to obtain actionable insights into user interactions. These insights guide the alignment of explanation usability with stakeholder needs. The overarching goal is to develop well-designed GUIs that enhance the accessibility and interpretation of the explanations provided towards comprehensible explanations that mitigate risks related to misinterpretation. For a successful implementation, this stage should cover user studies and address feedback from a diverse pool of test subjects.

4.2.4. QG Deployment

Model deployment, also referred to as model serving, scoring, or production, involves transitioning an evaluated AI model into operational use [DA22, 8]. This process ensures that the system is functional, operable, and compatible with other operational systems [ISO23c, 27]. Deployment includes installing, releasing, or configuring the model for operation in its target environment [ISO22a, 39]. Additionally, this stage involves preparing appropriate storage, handling, and shipping of systems that support operational readiness [ISO23c, 28]. It can be repeated multiple times to "[...] implement bug fixes and updates to the system" [ISO22a, 35], which is necessary, since failing to consider such deployment requirements early on may result in future problems [ISO22a, 37]. This phase is not focus of the present contribution. However, since the development and deployment stage are linked regarding identified information, we consider the extraction of relevant information, and information linking, when appropriate for our evaluation in part III, based on the *leaf-QG* information processing template.

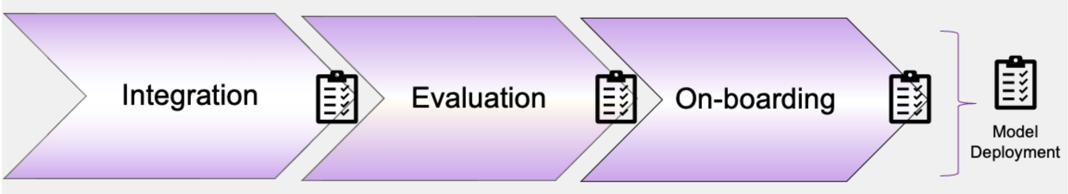


Figure 4.17.: *Collection-QG Deployment* comprises three sub-levels *Integration*, *Evaluation*, and *On-boarding*.

The proposed sub-processes, as depicted in Figure 4.17, are based on the CADAC lifecycle [DA22], and the *transition process* [ISO23c, 27] as defined in ISO 5338, whose purpose is to "[...] establish a capability for a system to provide the services as specified by stakeholder requirements in the operational environment" [ISO23c, 27]. In addition, the *deployment* phase lays the foundation for the *operation process* [ISO23c, 30], aiming to "[...] use the system to deliver its services" [ISO23c, 30]. *Operation* is further outlined as part of the *maintenance* phase, since *deployment* "[...] is smaller in scale [...]" [DA22, 8], and it "[...] typically involve[s] a smaller group of experts and users instead of organization-wide access" [DA22, 8]. Generally, deployment processes should enable the "[...] support [of] model updates [...] and the execution of continuous monitoring of established metrics associated with the use of the AI system" [ISO23c, 28], which may also involve on-boarding procedures to train users effectively. Figure 4.18 further outlines the proposed process structure for *deployment*, as introduced in the following.

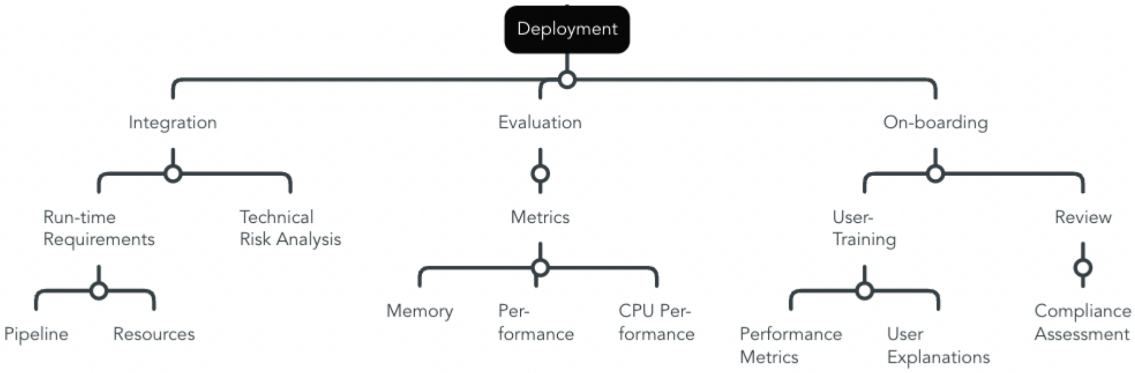


Figure 4.18.: Proposed compilation of processes and relevant information during the *deployment* stage.

4.2.4.1. QG Integration

Technical integration during model deployment addresses the operational requirements that often differ from the development environment [ISO23c, 28]. For instance, "[...] AI models can be deployed in a different format from how they were developed" [ISO23c, 28], as runtime requirements may necessitate alterations. Additionally, it may happen that "[...] runtime models are different from the models that are used in development, because the development environment is not compatible with runtime requirements" [ISO23c, 31]. Deployment modes can be tailored to operational needs. Models operate either in "[...] batch mode or in continuous mode, depending on whether the AI system has direct need for the model results" [ISO23c, 31]. Key considerations comprise the mode of use (real-time or batch processing), the number and types of end-users, the format and frequency of expected outputs, and the system's turnaround time [DA22, 8]. Furthermore, models may be deployed separately to accommodate specific hardware or software runtime requirements [ISO23c, 31], utilizing resources ranging from "[...] centralized cloud services and non-cloud data centres to servers (or clusters of servers), edge computing systems, mobile devices, and IoT devices" [ISO22a, 50].

4.2.4.2. QG Evaluation

Organizations must establish measures to "[...] assess how performance of the AI system can be affected after it has been put into use [...] and design appropriate monitoring metrics" [ISO23c, 28], building on the *development* evaluation stage where suitable metrics are identified. Documenting and defining these metrics is equally relevant to validation [ISO23c, 29], and continuous validation as part of the *maintenance* phase, which includes the definition of triggers for model updates [ISO23c, 29]. Regarding compliance assessment, for instance, a predetermined change control plan [PT23] documents expected changes, and how to handle them, including the definition of model update triggers [ISO23c, 28]. Further, with respect to verification, evaluating a successful deployment stage can include systems integration testing [ISO22a, 39]. Finally, during the deployment stage, technical risk classification and analysis must align with the preliminary risk assessment, considering the integration of the AI model with external systems and processes. This phase extends initial discussions of AI ethics, governance, and regulation, emphasizing their relevance post-deployment. The assessment should encompass all potential risks, including those affecting stakeholders, organizational functions, government regulations, societal norms, and broader societal implications. For instance, tools such as a risk register and risk assessment matrix can support evaluating the criticality of identified risks and for documenting appropriate mitigation strategies. [DA22, 8] The MQG4AI lifecycle blueprint is envisioned to offer a systematic approach that ensures comprehensive coverage and preparation for managing deployment-associated risks in alignment with individual AI systems as part of MQG4A, while providing access to comprehensive RAI design knowledge in form of MQG4DK.

4.2.4.3. QG On-boarding

To ensure a reliable system integration, the user may need additional training [DA22, 9]. For instance, this may include an introduction how to interpret the model's performance evaluation metrics and output, if they are visible in the final application, as well as how-to use the support system during *Maintenance*. This educational foundation is necessary for *human-machine teaming*, as outlined in section 3.3.2.1, which refers to the "integration of human interaction with machine intelligence capabilities" [ISO22a, 8]. Overall, during *deployment*, and "[s]pecifically in healthcare, further investigations in the form of observational studies, small-scale clinical trials, training, and user-acceptance exercises will be conducted" [DA22, 9]. Further, addressing legal considerations, an expert panel, steering committee, or regulatory body may conduct a comprehensive review of the project. This evaluation spans the dataset approach, AI model, evaluation metrics, and overall effectiveness. Additionally, processes for ensuring compliance, standardization, and post-implementation documentation, as well as the administration of contracts and service level agreements, are addressed. [DA22, 9] Finally, this stage also involves considerations for the legal protection of intellectual property, including patenting, trade secrecy for continued protection, or defensive publications in academic journals to advance the field [DA22, 9], as well as contributing RAI knowledge to MQG4DK could become an integral part of the deployment stage. Overall, the MQG4A planning structure for individual projects supports the review process, promoting responsible innovation.

4.2.5. QG Maintenance

Maintenance ensures that the system, which is "[...] running and generally available for use" [ISO22a, 39], maintains its capability [ISO23c, 31]. This stage is only broached within this contribution. However, since the development and maintenance stage are linked regarding identified information, for e.g. post-market monitoring, we consider the extraction of relevant information, as well as information linking, when appropriate for our evaluation in part III within the *leaf-QG* information processing template. As with traditional software development, maintenance in AI systems encompasses activities from previous lifecycle stages, which can evolve over time. Consequently, this phase builds on information extracted and concepts implemented during model development and deployment, as well as regarding the underlying data. AI models may require retraining or updates to adapt to changes in the world or shifting requirements. This involves tasks such as collecting new training data, modifying data preparation, and updating test data to ensure the system continues to function effectively. In some

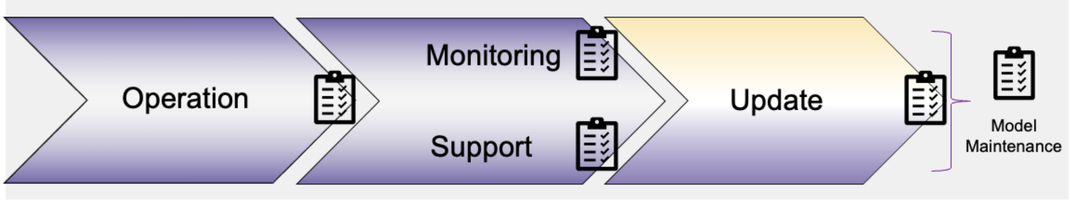


Figure 4.19.: Proposed sub-processes of *collection-QG Maintenance*. Monitoring the system's behavior and capabilities, as well as providing user support are illustrated in parallel, and model updates are closely related to the data stage, possibly triggering a restart of the complete lifecycle iteration.

cases, continuous training[9] replaces occasional retraining, enabling the system to maintain relevance and performance in dynamic environments. [ISO23c, 32] The proposed sub-processes (operation, monitoring, support, update), as depicted in Figure 4.19, align with the CADAC lifecycle [DA22], and the *maintenance process* [ISO23c, 31], including the previously, during *deployment* introduced *operation process* [ISO23c, 30], as defined in ISO 5338. In addition, *maintenance* comprises a *continuous validation process*, or monitoring of the system's performance and how it is impacted, as well as reasonable model updates, and possibly a model that continuously learns. [ISO23c, 32]

In summary, the maintenance of AI systems poses unique challenges due to their inherent complexity. An AI system's behavior can be unstable and not always explainable, even when supported by detailed engineering documentation. Additionally, the versions of configuration items may not reliably reflect the system's actual behavior. [ISO23c, 33] Finally, with respect to a responsible and long-term innovation integration, and leveraging the benefits of autonomous AI-driven processes, the question emerges, to what extent the AI system may operate in an automated manner.[10] With RM in mind, the following sections illustrate extracted maintenance sub-stages, as depicted in Figure 4.20.

4.2.5.1. QG Operation

During system operation, data for additional considerations is collected, regarding the consumed computing power and memory usage, which need to be mon-

[9] *Continuous learning* involves the model's ongoing evolution by incorporating production data into its training. This approach ensures the model adapts to shifts in desired behavior, reflecting changes in both input data and the corresponding outputs over time [ISO23c, 32].

[10] While the primary goal of AI is to support, enable, or enhance human actions and decision-making, hyperautomation goes a step further by fully automating these actions and decisions, all within the scope of ethical and regulatory requirements [DA22, 9]. This includes complex technical [DA22, 9], and ethical [Pal24] considerations, and, especially in the (medical) high-risk context, human agency and oversight is crucial, as outlined in section 3.3.5.1.

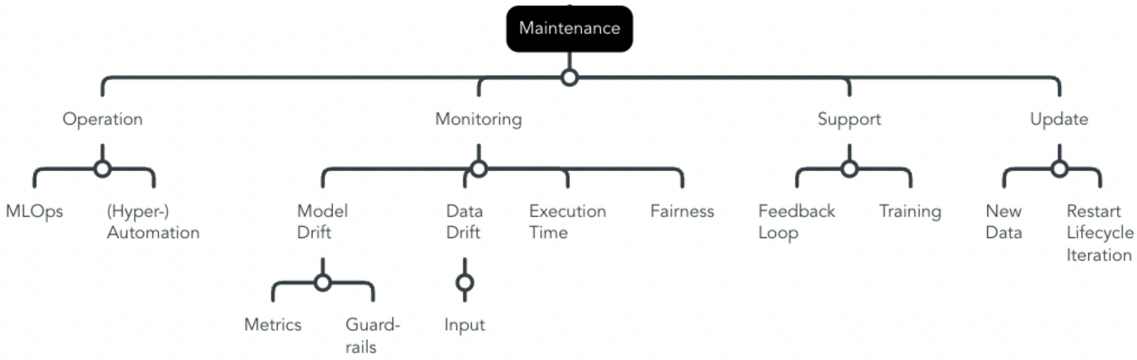


Figure 4.20.: Overview of proposed sub-levels of *collection-QG Maintenance*.

itored and evaluated, since they may impact further decision-making within the project, such as the frequency of training or the choice of the algorithm [ISO23c, 30], linking to the development stage. In addition, the successful operationalization of AI systems relies on AI pipelines, a process often referred to as AIOps or MLOps. This approach extends the well-established DevOps automation practices, tailoring them specifically for the deployment and ongoing management of AI models. [DA22, 9] Unlike DevOps, the concept of MLOps remains loosely defined due to the dynamic nature of the data input and the generated outputs of AI models. The operationalization process uses containers and microservices to build AI and data pipelines. Key aspects of the data pipeline include ensuring data availability, collection, storage, pre-processing, versioning, and addressing ethical considerations. Meanwhile, the AI pipeline must accommodate factors such as model compression, device compatibility, service definitions, versioning, auditing, re-training, maintenance, and monitoring. [DA22, 9] Despite advancements, most ML solutions have not yet been widely deployed in production settings, primarily due to their insufficient maturity compared to counterparts in other domains [QL24, 1]. Also, handling large volumes of dynamic data in projects often poses significant challenges [SF23, 1].

4.2.5.2. QG Monitoring

Monitoring how the performance of the AI system might change once it is in operation, is essential, and closely related with model *updates*. The evaluation of the AI system is guided by criteria that represent the technology itself, its application by diverse individuals in various settings, and the value generated by its use. These aspects form the basis for assessing the overall effectiveness and impact of the AI system. [DA22, 9] Generally, the AI system is monitored to ensure both normal operation and to detect incidents such as unavailability, runtime failures, or errors. These events are promptly reported to the relevant AI providers for corrective action. [ISO22a, 39]. By identifying potential factors that could affect performance, appropriate monitoring metrics are designed [ISO23c, 31]. As stated, "[a]n AI system, once deployed, can exhibit unexpected behaviour (e.g. as a result of bias or being exposed to unexpected input), and therefore monitoring performance is important, and procedures and processes can be more extensive compared to traditional systems" [ISO23c, 31]. Addressing contextual information on the real-world equally impacts the model development evaluation stage, which equally provides crucial information for RM. Performance monitoring corresponds to continuous validation, which ensures that "[...] AI models keep performing satisfactorily, or to demonstrate performance of the AI model over time" [ISO23c, 29]. Continuous monitoring procedures are crucial since the "[...] desired behavior can change" [ISO23c, 29], as well as the data encountered by the model [ISO23c, 29]. A strategy for executing post-market monitoring should be defined ahead of time [ISO23c, 32]. An example includes the previously, during deployment introduced, pre-determined change control plans for compliance review [PT23]. Overall, continuous monitoring of end-user activity is essential to evaluate how the AI model is performing within organizational functions. The level of activity observed can vary depending on the specific use case, with various metrics offering insights into the model's impact. These include user adoption, frequency of use, questions raised, the use or revision of documentation, feedback, and requests for new features. [DA22, 9] In addition, continuous validation plays a critical role in the ongoing improvement of RM processes, ensuring that the AI system adapts to evolving risks over time [ISO22a, 40]. Finally, QM plays a crucial role in overseeing the system's performance by documenting and assessing incidents, such as system failures and data errors, ensuring thorough tracking and timely responses [ISO23c, 31].

4.2.5.3. QG Support

Closely tied to system monitoring, is the option to provide users of the AI system with the necessary support to ensure their successful interaction with and utilization of the system [ISO22a, 40]. We propose treating user-related concepts as a distinct process that emphasizes the critical role of the user and provides room for the organization and monitoring of qualitative user studies. Generally, implementing effective support measures, such as feedback loops and user training, requires well-structured procedures, with all relevant information carefully collected in one place with shared multi-stakeholder-access, to ensure clarity and consistency.

4.2.5.4. QG Update

Intelligent systems are evolving and stochastic in nature, therefore updates are necessary. Overall, AI systems (software, models, and hardware) can be updated to address new requirements and enhance performance and reliability [ISO22a, 40]. These updates need to be scheduled in an organized manner, and in alignment with the previously introduced monitoring information. Therefore, the timing of updates depends on factors such as changes in operational processes, shifts in relevant data over time, observed declines in model precision, or the duration since the last update or model creation [ISO23c, 31]. The following monitoring directives for continuous validation are highlighted in ISO 5338 [ISO23c, 30], and they may result in different forms of model updates:

- Monitoring for data drift by checking if the model's input data deviates from what it was originally trained on. [ISO23c, 30]

- Monitoring for concept drift by assessing model performance on updated test data or identifying anomalies in output values, comparing recent outputs with previous ones. [ISO23c, 30] Model drift refers to the gradual decline in a model's accuracy caused by the evolving nature of data. This issue can often be mitigated by retraining the model using more recently gathered data that reflects these changes, thereby restoring its performance. [DA22, 9]

- Monitoring for other evolving requirements, such as changes in execution time, transparency, and fairness. [ISO23c, 30]

- In case of deviations, deciding whether to perform maintenance on the AI model. [ISO23c, 30] In contrast to model or concept drift, if model staleness is detected, the entire lifecycle is restarted. Model staleness arises when changes in the problem or environment that the model was designed for, lead to a mismatch between the model's assumptions and the current context. This can require a comprehensive review and adjustment of the model's architecture, inputs, algorithms, and parameters. [DA22, 9] This suboptimal performance, due to e.g. substantial differences in the production data, necessitates the execution of a rollback strategy for quick resolving [ISO23c, 32].

- Applying guard rails, if defined, to limit the model's output within certain boundaries or switching to a safer model if necessary. [ISO23c, 30]

- Finally, determining the frequency of validation to ensure timely and effective checks. [ISO23c, 30]

In addition, depending on the procedure, model updates may result in the disposal of (parts of) the AI system. This may comprise complex retrieval processes, depending on factors surrounding the intelligent system's real-world integration.

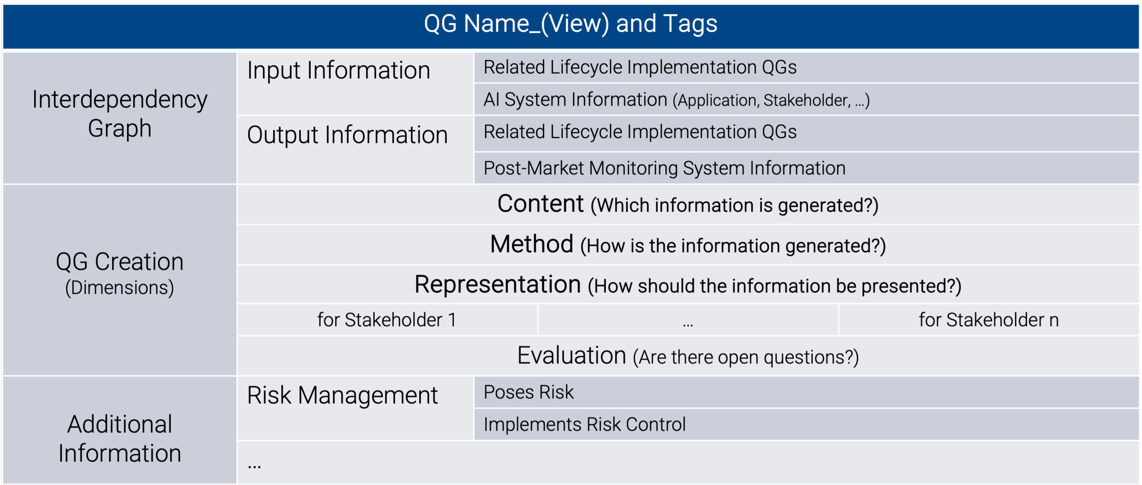
4.2.6. QG Decommissioning

Currently, we are at the early stages of integrating AI into real-world applications, and the *Decommissioning* phase is not the primary focus of our contribution. It is not always considered within existing publications that address the AI lifecycle [DA22] [QL24]. This area would benefit from more empirical experience, including an exploration of dependencies with other AI QM information. Overall, the *Decommissioning* phase outlines the essential process steps for withdrawing (components of) the intelligent system from its intended environment. This may encompass reversing the technical and human-level workflows that were previously established to ensure a seamless continuation of related processes in the real world. The disposal process, as defined in ISO 5338, aims to "[...] end the existence of a system element or system for a specified intended use, appropriately handle replaced or retired elements, and to properly attend to identified critical disposal needs [...]" [ISO23c, 33]. This includes a focus on data disposal, characterized as a "[...] new kind of disposal activity [...]" [ISO23c, 33], where addressing "[...] security or privacy risks [...]" [ISO23c, 33] is particularly critical when terminating data application. ISO 22989 takes a slightly different approach, framing disposal within the broader concept of retirement [ISO22a, 40]. This includes decommissioning and discarding the AI system and its data when the intended purpose is no longer valid [ISO22a, 40]. Additionally, ISO 22989 highlights the replacement of the AI system or components with more suitable alternatives, where "[...] repairs and updates are not good enough to meet new requirements [...]" [ISO22a, 40], as part of retirement [ISO22a, 40]. MQG4AI assigns the conceptualization of disposal processes to maintenance and model updates, leaning toward the ISO 5338 perspective. As a result, model updates may simultaneously trigger decommissioning processes, depending on the procedure, underscoring the interplay between maintenance, model updates, and the eventual disposal of (parts of) AI systems.

In summary, the lifecycle blueprint comprises a roadmap [GM09] for QG identification and positioning that incorporates RAI knowledge by design. After outlining MQG4AI's generic lifecycle information block in more detail in this section, the next section introduces the proposed *leaf-QG* information processing template towards reliable design decision-making.

4.3. The *Leaf-QG* Template – RAI Design Decision-Making

In addition to *collection-QGs*, *leaf-QGs* contribute a fundamental building block to construct the AI lifecycle within MQG4AI. They reflect concrete design decisions and provide an information processing layout, that aims for reliable design decision-making through considering AI's intricacies, including contextual information by design. Combined, these elements can be interpreted to derive interdependencies, forming a project-specific QG lifecycle graph structure. This structure incorporates (bi-)directional information connections, facilitating a comprehensive flow of data and dependencies throughout the lifecycle, including contextual information. Horizontal information on all levels of identified vertical connections is incorporated within individual *leaf-QGs* by default (input and output information).[11]



QG Name_(View) and Tags			
Interdependency Graph	Input Information	Related Lifecycle Implementation QGs	
		AI System Information (Application, Stakeholder, ...)	
	Output Information	Related Lifecycle Implementation QGs	
		Post-Market Monitoring System Information	
QG Creation (Dimensions)	Content (Which information is generated?)		
	Method (How is the information generated?)		
	Representation (How should the information be presented?)		
	for Stakeholder 1	...	for Stakeholder n
	Evaluation (Are there open questions?)		
Additional Information	Risk Management	Poses Risk	
		Implements Risk Control	
	...		

Figure 4.21.: The proposed *leaf-QG* template towards reliable design decision information processing.

Leaf-QGs are use case-specific within MQG4A, and address particular AI techniques within MQG4DK contributions, closing the gap from generic to targeted lifecycle design. The information layers, as depicted in Figure 4.21, are interpreted as quality criteria for design decision-making, which results in *leaf-QGs* executing a gate-keeping functionality within a designed scope.[12] Thanks to MQG4AI's generic and customizable character, information layers can be appended, removed, or declared as *optional*, tailored to the respective use case in a flexible manner. In addition, within MQG4A scenarios, evaluating to what degree the combined leaf-QGs are filled, may serve as foundation for comprehensive quality evaluations within the QG scoring system.

Overall, we aim for sufficient generalizability, so that the proposed information structure is applicable to as many lifecycle stages, AI techniques, use cases, and

[11] Depending on the hierarchical level of the vertical *collection-QG* graph structure, they could be equipped with information layers to contribute to a more high-level interdependency design, and equally extract information on horizontal connections with other QGs.

[12] *Collection-QGs* may equally function as gatekeeper, if, for instance, they point out necessary process steps, such as the definition of model evaluation procedures.

design decisions, as possible. The current template layout is mainly based on experiences during model development, related data utilization steps, and, highlighting *Interdependency Analysis*, considers information extraction with respect to deployment and maintenance, including relevant contextual information. However, different phases may require a different approach. In general, aiming towards compliance, *leaf-QGs* are intended to "[...] capture in precise terms the processes, techniques and methods needed to make AI systems trustworthy in a verifiable manner [through organizing particular design decisions], ensuring they address all identified risks in line with the Regulation, while being mindful of the implementation burden" [SG24b, 3]. The *leaf-QG* template is demonstrated in detail in chapter 6, through adaptation to a technical guideline for evaluating the quality of LIME and SHAP explanations [LGC24] as a contribution to MQG4DK. This section illustrates the identified *leaf-QG* layout, before outlining our medical use cases that comprise the foundation for MQG4AI's evaluation in chapter 5, next.

4.3.1. Individual *Leaf-QG* Creation

The aim of *leaf-QGs* is to illuminate the underlying (continuous) process of defining and monitoring individual design decisions as a foundation for achieving quality by design, contributing to responsible lifecycle planning. Therefore, *leaf-QGs* provide an information processing structure to guide contributing stakeholders. This section first introduces the concept of *Guiding Questions* (GQ) for information extraction based on the Socratic method [IA12], before outlining the proposed *leaf-QG* information dimensions for responsible design decision-making.

4.3.1.1. Guiding Questions & the Socratic Method

Overall, QG Creation refers to the process of systematically identifying relevant aspects to ensure the trustworthiness of AI systems through shaping the individual design decision-making process. One possible method to support this information extraction is the use of *Guiding Questions*, inspired by the *Socratic Method* [IA12]. The *leaf-QG* visualization on GitHub[13] provides an example for generic GQs, and is based on Figure 4.21.

Socratic Method The Socratic Method is a teaching technique based on guided question–answer dialogues that help learners critically examine their existing knowledge and reasoning, rather than simply evaluating their responses. It involves a two-step process: first, the teacher uses questions to reveal gaps or flaws

[13] https://github.com/miriamelia/MQG4AI/blob/main/templates/Template_LeafQG.md

in the learner's current understanding, creating awareness of the need for deeper inquiry. In the second step, the teacher acts as a facilitator rather than an authority to support learners in discovering new knowledge by reflecting more deeply on the issue. This method shifts responsibility to the learner and encourages active, self-driven learning. [IA12, 358] Within MQG4AI, the 'teacher' corresponds to the *leaf-QG* template, and information layers that are accompanied by GQs, are intended to result in responsible design decisions. For the second step, depending on the individual use case, possibly other stakeholders from different backgrounds may identify relevant questions to further evaluate the design decision's implementation in an iterative manner during regular meetings, for instance. The identified questions and information, in turn, are envisioned to be documented within MQG4DK to grow the RAI knowledge base, as intended by DSR [vJ20], which is detailed in section 3.3.7. A generic GQ aimed at identifying relevant *Output Information* of the respective design decision could be summarized as follows: *What information is produced that is relevant to other stages and design decisions?* Another, slightly more specific example for design decisions that are related to explaining model output may result in: *What assumptions does the model make about the input data, and how could these affect different user groups?*

Generally, the questions are intended to encourage critical reflection and structured thinking to uncover important information across various use cases [HY23]. Socratic questioning is even adapted to prompt LLMs, displaying promising results in complex reasoning tasks [QJ23]. Additionally, "[...] this method promotes motivation to enhance better learning [...]" [IA12, 357] and participating stakeholders are "[...] more autonomous and [take] control of their learning by discovering knowledge in dialectical dialogs" [IA12, 360]. Last but not least, "[i]t facilitates the construction of knowledge through discourse based on personal experience and this can create a culture of knowledge sharing" [RG09, 1], which contributes to enhancing KM within the MQG4AI lifecycle blueprint, as well as AI literacy in general. In summary, the Socratic Method represents a potentially valuable instrument for stakeholders involved in defining design decisions. The foundational leaf-QG template presented in this section, serves as an initial framework to support systematic exploration of relevant considerations. The continued design and empirical evaluation of GQs constitute essential components of future work.

4.3.1.2. Three Core Dimensions

The proposed leaf-QG information extraction is grounded in a threefold structure: *Content* definition, *Method* outline, and stakeholder-tailored content *Representation* of relevant information.[14] In addition, the proposed information pro-

[14] We focus on extracting information for the *Content* and *Method* dimensions during development, and only broach *Representation* for MQG4AI evaluation in part III.

cessing structure aims to facilitate testing by integrating multi-perspective transformations into the design of acceptable implementation approaches through the *Representation* layer. Finally, an *Evaluation* layer provides room to document possible gaps within the chosen method for future reference.

Dimension 1: Content Definition The *Content* dimension delves into the design decision in greater detail, addressing the question: *"What information is generated?"* Aiming to be sufficiently generic for application across all lifecycle design decisions, it has been evaluated from multile perspectives to provide a comprehensive information overview for the development stage. Concretely, we illustrate contributions to MQG4DK for a performance evaluation metrics *leaf-QG* compilation (chapter 7), a single *leaf-QG* surrounding a technical guideline for assessing explanations (chapter 6), and criteria for segmentation model selection as part of simulating MQG4A-template versions (chapter 8), which can be viewed on GitHub[15]. Overall, in alignment with the European Commission's statement on AI standard development, we aim for "[s]ufficiently prescriptive and clear" [SG24b, 3] guidelines for design decision-making. *Leaf-QG* contributions are intended to provide regulators, and developers with a hands-on approach to the respective design choice. MQG4A-scenarios include concrete results, while MQG4DK contributions may be based on a fictional use case for better contribution of more use case-specific RAI design knowledge.

Dimension 2: Content Definition Method The *Method* dimension describes the process of making design decisions, answering the global question *"How is the information generated?"* The provision of detailed guidance on how to implement the respective QG is necessary to ensure a solid concept, that avoids AI pitfalls [HSA22] [TJ21] [MF22], while responding to the technique's evolutionary character. In addition, outlining the underlying method in more detail, serves to evaluate the design choice's reliability, and thus contributes to compliance assessment and QM. Equally oriented towards the setup of standards, monitoring the underlying method in addition to the implemented design decision, fulfills specifying "[...] the key technical methods and techniques that support [unique aspects of data-driven AI systems] in practice, including criteria and priorities to ob-serve when defining and measuring these properties for specific AI systems" [SG24b, 4].

Dimension 3: Content Representation *Representation*, the final dimension, focuses on determining which information should be communicated to specific stakeholders (1..*n*). Generally, depending on the knowledge and background of the respective stakeholder who is the recipient of information on a concrete design decision, knowledge may need to be presented in a different format. For instance, a developer consults generated model explanations with a different mindset that a user, or project manager. This information layer summarizes considerations on the appropriate timing, as well as the transformation of content to meet

[15] `https://github.com/miriamelia/MQG4AI/blob/main/README.md`

the needs of identified stakeholders: *"How should which information should be presented to which stakeholders and when?"* Refer to sections 3.3.2.1 and 3.3.3.5 for more information on stakeholder inclusion and prominent roles.

Evaluation Finally, the *Evaluation* layer identifies open questions and potential limitations of the chosen approach for implementing a design decision: *What are open questions when applying the generated information?* This layer may equally serve as a basis for conducting (residual) risk analysis through assessment of the chosen method. In addition, highlighting the iterative character of AI lifecycle design, identified follow-up tasks that are not (yet) included within the *Method* dimension may be documented here.

As a result of these guided information transformations surrounding individual design decisions, contributing stakeholders are required to reconsider information by design, which supports a solid concept definition. In addition, the *leaf-QG* information structure enables "[...] the participation of representatives from all sectors and types of organisations [...]" [SG24b, 2] within the AI project (MQG4A), as required for European AI standards [SG24b]. Further, MQG4DK contributions that address specific stakeholders, their roles, and their relation with AI, for instance, can be appended for shared design knowledge.

4.3.1.3. Horizontal Interdependency Graph

The *Interdependency Graph* addresses AI's inherent intricacies through the identification of relevant input and output information of design decision-making towards risk mitigation by design. The relevance of *Interdependency Analysis* is further outlined in section 3.3.2.3. This *leaf-QG* information layer incorporates related information drawn from other lifecycle stages and supplementary information blocks, emphasizing their interconnectedness. An example to align a use case- and model-agnostic approach to evaluate explanations with the *leaf-QG* template is illustrated on GitHub[16]. Overall, this setup is envisioned to facilitate a comprehensive analysis of interdependencies through the provision of a holistic perspective on the broader scope of lifecycle QGs.

Input Information The *Input Information* layer focuses on the question: *What information is required to execute the method and produce the content?* This includes relevant AI system details, such as the intended use or necessary stakeholder roles, as well as information from other lifecycle QGs that support the creation of individual QGs, highlighting the role of data, for instance.

[16] `https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/2_Lifecycle/2_Development/4_Model_Explanation/Method_Evaluation/Quality/QG_FidelityRobustnessScore_(SHAPLIME).md`

Output Information The *Output Information* layer addresses the question: *Which stages are impacted, and what additional information is extracted?* It emphasizes identifying and linking relevant information for post-market monitoring during maintenance, for instance, while establishing connections with other lifecycle QGs. Overall, integrating knowledge from and with related information blocks aims to ensure comprehensive system understanding and alignment in a continuous manner.

4.3.1.4. Additional Information Layers

Concluding the introduction to the leaf-QG template, supplementary information layers close the gap to comprehensive IM for RAI applications, demonstrated for TAI RM in Figure 4.21. The format is adaptable to various stages and types of information, and the customizable template is designed to scale and incorporate AI QMS requirements through information linking that may communicate with an external RAI information block, or collect valuable information for system-wide evaluation mechanisms.

Risk Management Layer Information linking is illustrated for RM within our contribution, and the *Risk Management layer* connects individual design decisions to identified risks, facilitating the ongoing implementation of risk controls and iterative (residual) risk analysis throughout successive MQG4A-versions, and possibly following MQG4DK contributions. This includes addressing potential risks associated with each design decision, by first identifying any related risks (*"Are there any associated risks with this design decision?"*) and then outlining the risk controls implemented to mitigate them (*"What risk controls are put in place to mitigate these risks?"*). This setup, within the broader picture of MQG4AI that identifies risks based on TAI, is aligned with requirements for AI standards. They "[...] should be aimed at identifying and mitigating risks of AI systems on the health, safety and fundamental rights of individuals. This is a novel aspect for AI standardisation, as the orientation of published and ongoing ISO/IEC work takes a very different ap-proach in terms of risk objectives and definitions" [SG24b, 4].

Outlook: Customizable Layers Finally, customizable information extraction layers can be appended, for instance aiming to fulfill further AI QMS requirements, as imposed by the AI Act in Article 17 [Fut24], as discussed in section 3.4.1.2. For instance, this may include a resource layer that links the required GPU-usage with concrete design decisions. In addition, this may comprise information layers that tackle more implementation-oriented information on utilized software, packages, and, in addition to the identified input and output dependencies, possibly a data layer could be reasonable. For instance, ISO 25059 on *Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems* [ISO23a], offers information on further information layer extraction from a software engineering viewpoint tailored to AI. How it relates with other ISO quality standards is introduced in detail in [OJ24]. Overall, the suitable setup of different information layers needs to be tested with further, and more complex MQG4A-scenarios that span entire projects, which is beyond the scope of this thesis.

After outlining the thematic dimensions that comprise the *leaf-QG* information extraction layers for design decision-making, the next section introduces MQG4AI's *leaf-QG* modules that are intended to enable the implementation of decentralized knowledge fusion. This integration of DSR [vJ20] towards comprehensive and continuous lifecycle planning results in intelligent systems that benefit from shared RAI knowledge.

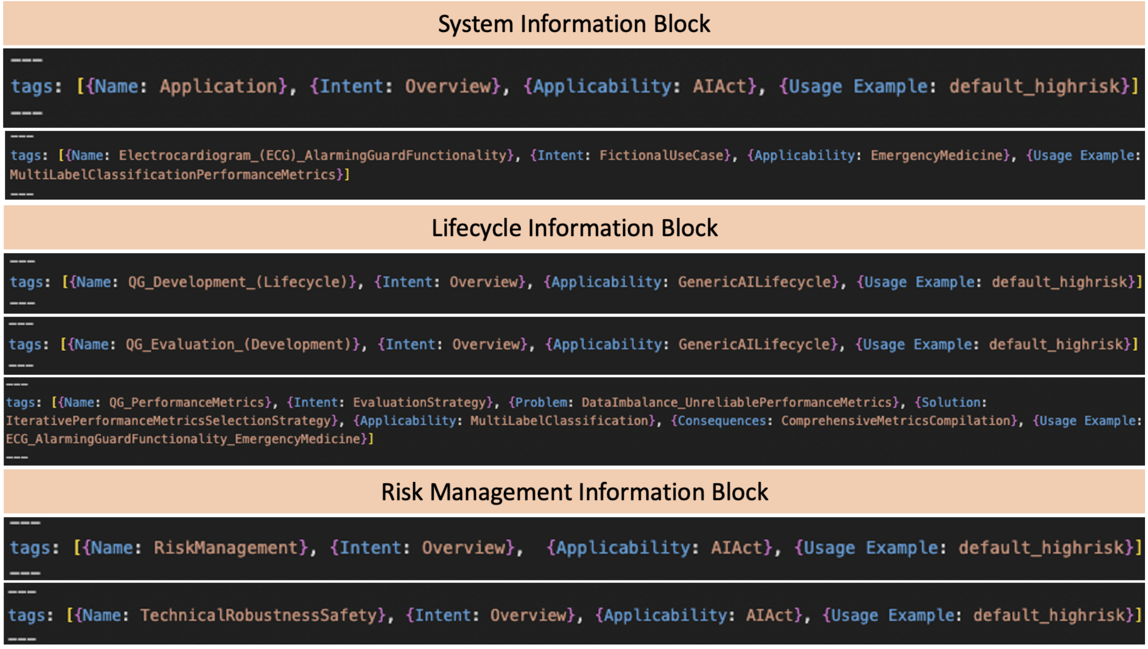
4.3.2. Decentralized Knowledge Fusion (MQG4DK & MQG4A)

As introduced in sections 3.3.7.1 and 4.2.1, the use case adaptation of MQG4A is intended to be implemented through a configurable pull version from MQG4DK (MQG4A-v0). Therefore, the *leaf-QG* format incorporates information layers that facilitate intelligent searching based on use case-specific attributes through *leaf-QG-tags*. This ensures that only relevant lifecycle design decisions are retrieved from the design knowledge base, along with the generic default template, which is based on the living lifecycle blueprint for high-risk systems to generate use case-specific template versions. In addition, *leaf-QG-naming* contributes to AI knowledge organization within MQG4AI.

Leaf-QG Naming The *QG Naming*-structure is intended to link information with AI techniques that share a particular structural setting, such as classification or segmentation scenarios. This should be align with a reasonable use case classification approach as part of future work. Additionally, QG naming supports the identification of horizontal lifecycle interdependencies, highlighting shared design-decision contributions to MQG4DK in form of multiple leaf-QGs.

Example For instance, the design decision *Reliable Performance Evaluation Metrics* has different implementations depending on the use case, and consists of multiple related QGs along the lifecycle, such as the inclusion of additional material for metrics analysis. This results in a combination of related QGs for reliable metrics, including *QG_PerformanceMetricsCompilation_(Classification)*, and *QG_ConfusionMatrix_(ClassificationPerformanceMetrics)* for classification scenarios, as outlined in chapter 7. Another example of our contribution addresses the lifecycle collection-QG explanation of the model development stage, which includes approaches to evaluate the quality of explanations. Different implementations depending on the chosen XAI method are possible. This results in multiple potential leaf-QG contributions to MQG4DK that fulfill the same or objective in a different manner. Consequently, the fidelity-robustness score [LGC24], which is introduced in chapter 6, is referred to as: *QG_FidelityRobustnessScore_(SHAPLIME)*.

Leaf-QG Tags *QG Tags* are essential for identifying use case-specific lifecycle stages that allow for MQG4A to be pulled from MQG4DK based on a configuration file and an intelligent tag search. Supplementary information blocks, along with the identified high-level lifecycle phases, are included in the default high-risk MQG4AI lifecycle blueprint. Consequently, they are represented in MQG4A-v0, and characterized by the tag **Overview**, as shown in Figure 4.22. The proposed information structure for tagging lifecycle stages is based on design patterns for ML applications, as outlined in [WH22]: *Name, Intent, Problem, Solution, Applicability, Consequences, Usage Example*. Other approaches to characterize ML design pattern may also be viable, and their evaluation will be addressed in future work with more practical experience.



System Information Block

```
---
tags: [{Name: Application}, {Intent: Overview}, {Applicability: AIAct}, {Usage Example: default_highrisk}]
---
```

```
---
tags: [{Name: Electrocardiogram_(ECG)_AlarmingGuardFunctionality}, {Intent: FictionalUseCase}, {Applicability: EmergencyMedicine}, {Usage Example:
MultiLabelClassificationPerformanceMetrics}]
---
```

Lifecycle Information Block

```
---
tags: [{Name: QG_Development_(Lifecycle)}, {Intent: Overview}, {Applicability: GenericAILifecycle}, {Usage Example: default_highrisk}]
---
```

```
---
tags: [{Name: QG_Evaluation_(Development)}, {Intent: Overview}, {Applicability: GenericAILifecycle}, {Usage Example: default_highrisk}]
---
```

```
---
tags: [{Name: QG_PerformanceMetrics}, {Intent: EvaluationStrategy}, {Problem: DataImbalance_UnreliablePerformanceMetrics}, {Solution:
IterativePerformanceMetricsSelectionStrategy}, {Applicability: MultiLabelClassification}, {Consequences: ComprehensiveMetricsCompilation}, {Usage Example:
ECG_AlarmingGuardFunctionality_EmergencyMedicine}]
---
```

Risk Management Information Block

```
---
tags: [{Name: RiskManagement}, {Intent: Overview},  {Applicability: AIAct}, {Usage Example: default_highrisk}]
---
```

```
---
tags: [{Name: TechnicalRobustnessSafety}, {Intent: Overview}, {Applicability: AIAct}, {Usage Example: default_highrisk}]
---
```

Figure 4.22.: Examples for *leaf-QG-tags* within MQG4AI's information blocks, reflecting different hierarchical levels of MQG4DK contributions.

In summary, the, in this section introduced generic and customizable *leaf-QG* information processing format is designed to enable a decentralized fusion of collective knowledge on RAI lifecycle implementations, aiming to cover the variety of possible AI use cases. This is achieved through the ongoing interaction between abstract MQG4DK and applied MQG4A along the AI lifecycle. MQG4DK contributions can be adapted to a specific AI use case through an intelligent, tag-based search. MQG4A lifecycle design elements are created and updated during the conceptualization stage to organize, implement, and continuously monitor relevant design decisions in a responsible manner. Filling in the operational information layers (QG tags, QG naming) could possibly be automated when implementing MQG4AI as a tool. In the long term, the MQG4AI lifecycle planing blueprint is designed to improve the selection of existing methods that mitigate AI risks, helping to prevent potential issues while documenting the lifecycle concepts for all contributing stakeholders, and under the provision of shared access. This allows for a comprehensive and continuous planning phase. Finally, a key premise of the proposed approach is that AI use cases are reasonably categorized, e.g. based on structural similarities and domain-specific factors, in addition to viewing the AI system as a combination of its underlying lifecycle design decisions.

After detailing QGs and lifecycle design within MQG4AI in this chapter, the next chapter 5 introduces the medical domain, particular challenges of intelligent applications in healthcare, as well as our two very different use cases, that form the foundation for MQG4AI evaluation in part III. Namely, we analyze intelligent ECG classification in emergency medicine (EM), and intelligent enhancements of a medical software surrounding *Achalasia*, a rare disease of the esophagus.

5 Application Domain Medicine

AI in medicine holds immense promise but also presents a unique set of challenges. The medical domain is inherently complex, multidimensional, and deeply intertwined with both the human body and mind, making it a fascinating yet demanding field for AI integration. We chose use cases situated in this high-risk domain to design MQG4AI, as introduced earlier in part II, and evaluate interaction scenarios in part III. Therefore, this chapter first introduces RAI challenges in medicine. In addition to addressing complex domain knowledge and data considerations, ethical aspects are also discussed. Further, from a regulatory perspective, the high-risk nature of medical applications arises from their direct impact on human lives, placing them under strict oversight within a multilayered regulatory system that includes GDPR, MDR/IVDR, and the AI Act, as introduced in section 3.1.

Overall, ensuring compliance while harnessing AI's transformative potential remains a critical balancing act, which we aim to support with MQG4AI. Therefore, in addition to outlining consideration surrounding medical AI, this chapter introduces our two use cases that comprise the foundation for evaluating MQG4AI interaction scenarios in part III. Namely, intelligent electrocardiogram (ECG) classification in emergency medicine (EM), as well as a medical software tailored to the rare disease Achalasia, a motility disorder of the esophagus.

5.1. Challenges of AI in Medicine

AI's inherent dynamics are interpreted in healthcare, and "[...] it is crucial to define the categories of use cases for [AI] in the clinical setting [...]" [ISO21b, V]. ISO/TR 24291:2021(E) on *Applications of machine learning technologies in imaging and other medical applications* provides a comprehensive overview to identify and classify intelligent medical applications, broadly organized into three categories: *technology* (e.g. AI techniques such as DL, robotics, or NLP), *medical specialty* (e.g. cardiology, emergency, or dermatology),[1] and *medical usage* (e.g. for clinical trials,

[1] Note that rare diseases are not listed as a medical specialty.

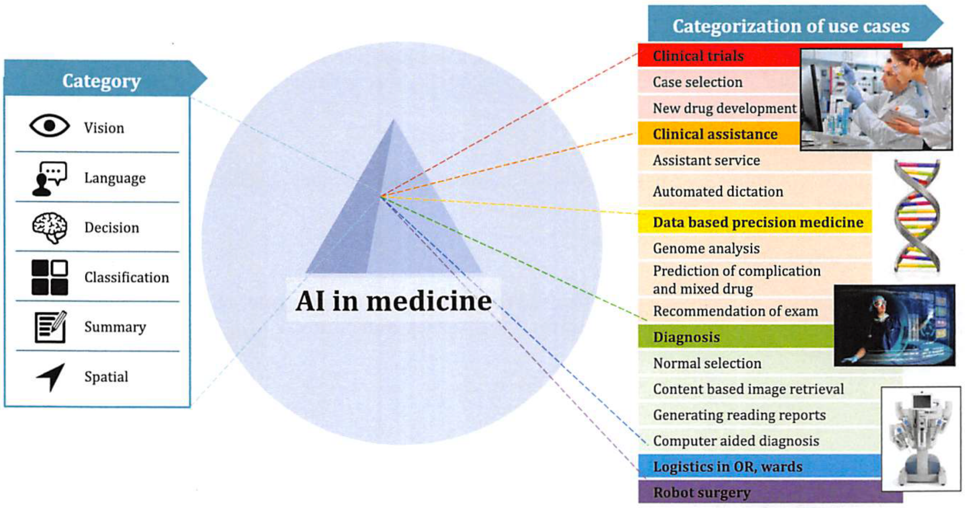


Figure 5.1.: Use case categorization of medical AI, as depicted in ISO/TR 24291:2021 [ISO21b, 10], highlighting medical usage.

assistance, or robot surgery) [ISO21b, iii]. This structure provides a solid starting point to identify scenarios for novel contributions, while providing an overview on the current state of AI in medicine [ISO21b, V]. Generally, model quality is impacted in use case-specific ways, and "[...] all AI tools are prone to errors" [ES24, 1]. Therefore, a profound and comprehensive domain analysis in alignment with technical design choices enhances model quality. Further, a comprehensive report, published by the European Commission, emphasizes the social impact of AI in medicine, highlighting ethical questions [GGG20]. For instance, they depict a 'visual overview' of AI in medicine, including a social impact estimation, reaching from *negative* (e.g. bio-terrorism, the quest for immortality, or the search for artificial life forms), over *controversial* (e.g. genetic tests, brain-machine interfaces, or precision medicine) to *positive* (e.g. computer-aided diagnosis, eHealth, or precision medicine) [GGG20, 27].[2]

This section provides an overview of a selection of AI challenges in medicine, aiming to capture technical, ethical, as well as regulatory challenges towards implementing RAI. Namely, we highlight the role of complex medical domain knowledge, the human influence along the AI lifecycle, interpretation and handling of AI outputs by the user, as well as the importance of discussing ethical questions, all of which result in the need for diverse teams. In addition, medical data, that usually is scarce compared to other industries, has its own particularities regarding different formats, constituency, and quality, for instance. Finally,

[2] Note that precision medicine, i.e. individual patient-tailored methods and outcomes, is interpreted as having *negative* and *positive* social impacts, depending on the respective application mode. Tailored treatments and drug design bear immense potential to enhance global healthcare, while individual profiling may result in risks affecting AI trustworthiness that need to be addressed regarding e.g. data security and privacy. [GGG20, 27]

complementing technical and ethical challenges, healthcare is characterized by a complex and multi-dimensional regulatory landscape, see section 3.1.

5.1.1. Intricate & Dynamic Medical Domain

Complex domain knowledge impacts design decision-making in various ways along the AI lifecycle [PJ21b] [EM24] [MD23]. For instance, (performance) evaluation is closely tied to the individual use case, since the respective medical domain impacts design decisions through influencing tuning objective(s), and choice of metrics compilation, among others. In addition, the interpretation of how to integrate AI output with existing real world workflows, results in additional challenges. For instance, domain-embedded risks such as the "reinforcement of outdated practices" [ES24, 1] through information provided by the AI, need to be addressed during conceptualization in a continuous manner. Therefore, project success equally necessitates the establishment of communication channels early on, which is crucial to enable a diverse and interdisciplinary AI lifecycle approach in alignment with the real world's evolutions in healthcare.

Section 3.3.2.2 emphasizes *Domain Knowledge*, addressing AI's use case specificity, and section 3.3.2.1 on *Stakeholder Inclusion* further outlines the need to translate domain knowledge both ways, domain expert knowledge for developers, and vice versa regarding contributing stakeholders, for instance. Finally, considering the human user plays a critical role in healthcare, however, "[...] little research has been devoted to real world translations with a user-centered design approach" [AC21, 1].

5.1.2. Human-Centered Design

The relationship between the "algorithm and users" [HC21, 1] is especially relevant in healthcare. Medical users comprise medical personnel, and/or patients, for instance. Therefore, questions surrounding balancing model opacity and explainability are crucial topics for medical AI, resulting in the need for research and implementation of human-centered methods for AI system creation, towards human-machine teaming [Hen22]. However, human-centered design can be "[...] challenging due to the limited access to end users and the knowledge imbalance between those users and ML designers" [HC21, 1]. An important related topic, for instance, comprises *automation bias* [ES24, 1], or when medical users trust the AI's output without second-guessing. Such trends should be considered when designing the GUI, embedded within the respective domain. Additional topics may include the provision of information on data bias. For instance, if the system was trained on mainly *male* data samples, and it encounters a *female* patient, information on the training data's impact should be communicated in a reasonable manner to the user, highlighting model accuracy and confidence in its output, among other criteria. As a result, "[h]uman factors researchers should actively be part of efforts in AI design and implementation [...]" [AC21, 1] to conduct "formative user research" [HC21, 1], aiming for a successful communication of relevant information. Generally, whether or not the model is continuously learning, or a next updated version is already in training, the current state(s) of the model in production need(s) to be "screenshoted" to extract relevant information that is translated and communicated to relevant endpoints, reaching other stakeholders (or machines) in an on-going manner.

5.1.3. Ethical Questions

Generally, regarding the medical domain in particular, lots of ethical questions emerge. They comprise topics addressing bias [EM25b] and how to handle different sources and forms towards fair outcomes, which is highly use case-specific and real world dependent. Information on AI ethics and social impact is provided by the European Commission, and for instance, [GGG20, 28] demonstrates concrete examples for ethical trade-off discussions. For instance, they comprise the degree of automation and human-in-the-loop-approaches, aiming to "[...] capture the complexity and diversity of real-life applications [...]" [Pal24, 1]. The social impact of fully autonomous AI systems, such as the *digital doctor*, or *robotic surgeon* are estimated to result in a rather negative outcomes [GGG20, 27]. Other perspectives demonstrate a different opinion, highlighting advances of automation and related opportunity costs when pursuing a more risk-averse stance: "[...] there are some scenarios where an insistence on keeping humans in the loop (or in other words, the resistance to automation) seems unwarranted and could possibly lead us to miss out on very real and important opportunities in healthcare — particularly in low-resource settings" [MK24, 1]. Further, "[...] where certain criteria hold (e.g. the AI is as accurate or better than human experts, risks are low in the event of an error, the gain in wellbeing is significant, and the task being automated is not essentially or importantly human), it is both morally permissible and even desirable to kick the humans out of the loop" [MK24, 1].

Finally, both perspectives result in important considerations. Establishing a balance for the individual project within its context is crucial. A possible method of balancing both views, lies in enhancing the quality of the human-in-the-loop. For instance, this can be implemented through a comprehensive summary and efficient schedule grounded in research and education. As a result, the responsible person is educated and self-sufficiently equipped for ethical AI decision-making. Human and data quality are interrelated. Towards long-term RAI, both sources need to be addressed in a holistic way within their real world setting.

5.1.4. Medical Data

Data is essential to grow AI models, it provides an access point to the real world. But, medical data is especially expensive, diverse, and complex to preprocess, interpret, as well as label. The following list summarizes data-related challenges in medicine, based on [ES24, 281]:

- *Support* ML data samples usually are divided into *features* (i.e. the pool of information the model aims to interpret) and *labels* (i.e. the target data the model trains towards and aims to connect with feature patterns). For ML to be successful, a sufficient amount of data is crucial to find patterns and learn. However, due to various reasons, including high labeling costs, or complex data acquisition of medical raw data, healthcare is characterized by data scarcity. Especially, regarding rare diseases, that necessitate more medical research and are insignificantly represented, statistically speaking. This data imbalance comprises a fundamental source for bias, which can result in unfairness and harm patient well-being. Therefore, continuously monitoring the data distribution is of essence.

- *Formats* A multitude of medical data formats exists, relevant types of which depend strongly on the particular use case. For instance, in medical imaging, *Digital Imaging and Communications in Medicine* (DICOM)[3] files are used for ML with e.g. CT, ultrasound or X-ray scans, aiming towards interoperability between devices. Or, *Fast Healthcare Interoperability Resources* (FHIR)[4] comprises a standardized data concept "[...] for exchanging health care information electronically." This diversity results in complex infrastructures and, related to AI, computing costs. In addition, many medical conditions are complex in nature, leading to multi-dimensional input and output data.

- *Costs* Further, investments are needed due to e.g. high infrastructure, regulatory, and labeling costs. Regarding the latter for instance, a successful integration of AI necessitates qualified personnel. Possibly, more focus on user-trainings, aiming to transfer complex knowledge to a group of people, provide a means towards long-term well-doing projects, which could mitigate costs.

- *Variance* Generally, medical raw data is characterized by high variance in data quality, which necessitates comprehensive data preparation for ML. In addition to technical methods, such as *normalization*, or *clipping*, for instance, different resolutions are based on a variant data acquisition process. Depending on the medical human-in-the-loop who collects the data, the concrete acquisition is carried out slightly different each time. Therefore,

[3] `https://www.dicomstandard.org/standards/view/introduction-overview`

[4] `https://ecqi.healthit.gov/fhir`

establishing standards for medical data acquisition is expected to have beneficial effects, among other ideas.[5]

- *Translationality* Finally, the creation of solid data infrastructures, intra-hospital, inter-hospital, on national and European level is challenging due to a multitude of reasons. But necessary to continuously build strong AI. In addition to establishing new infrastructures, such as recently emerging Electronic Health Records (EHR) in Europe,[6] that emphasize secure access, horizontal and sector-specific legislation plays a crucial role when shaping AI's arrival in the real world.

5.1.5. Multi-dimensional Regulatory Landscape

Focusing on the AI Act, medical regulations, and data protection under the GDPR, as discussed in section 3.1.1, a complex and evolving multi-dimensional regulatory landscape emerges. To bridge the translational gap toward effective RAI implementation, outlined in section 3.3.5.7, the clear allocation of liability and responsibility is particularly critical for intelligent medical applications. This involves several levels of execution. For example, providers are required to implement an AI QMS in collaboration with a notified body. Additionally, government-led initiatives support the creation of a broader compliance ecosystem, which includes infrastructures such as regulatory sandboxes (as defined in Article 57 of the AI Act [Fut24]) and tiered communication channels designed to enhance AI literacy and stakeholder engagement. Currently, established guidelines are missing for implementation. The EU AI Office[7] is tasked with implementing an AI Act-conform ecosystem in Europe, and to provide guidance to not yet fully answered questions. Implementation processes are in their infancy, which includes the development of standards that address concrete AI-related challenges (in medicine). Among others, methods that implement continuous monitoring in a reliable manner, as well as AI RM are highlighted.

In addition, with respect to requirements summarized in the MDR/IVDR, a double-regulatory burden may emerge, which is not intended. Alignment is an on-going discussion, and subtle differences emerge e.g. regarding differently valued concepts, such as emphasis on the, in section 3.2.1.2 introduced risk-benefit ratio approach during RM in medicine, which is not explicitly mentioned in the AI Act. In addition, ethical healthcare-related questions consider debates whether AI use needs to be communicated to patients, in an obligatory manner. Continuing, for instance, research exemption is treated differently in both regulations. In Recital 25, the AI Act states that "[t]his Regulation should support in-

[5] `https://www.cdisc.org/standards/foundational/cdash`

[6] `https://digital-strategy.ec.europa.eu/en/policies/electronic-health-records`

[7] `https://digital-strategy.ec.europa.eu/en/policies/ai-office`

novation, should respect freedom of science, and should not undermine research and development activity. It is therefore necessary to exclude from its scope AI systems and models specifically developed and put into service for the sole purpose of scientific research and development" [Fut24]. However, according to the MDR, non-commercial projects need approval by the ethics commission. Overall, the resulting tradeoff between regulation and innovation, and with respect to medicine, fear of over-regulation is a continuous discussion with urgent need for action to ensure a high standard of healthcare in Europe. Finally, this process is carried out by a multitude of different human beings, and the need for qualified personnel is omnipresent.

In summary, the big question *how to implement and prove compliance for high-risk AI (in medicine)* accompanies AI's arrival in the world. Key topics comprise transparency, human oversight, accuracy, robustness or cybersecurity (Articles 13-14-15-16 of the AI Act [Fut24]). These considerations are rooted in existing ethical questions that need to be tailored to the respective application scenario. In medicine, patient well-being is the most important directive. After outlining particular challenges surrounding AI in the healthcare domain, the next sections introduce two very different medical use cases that comprise the foundation to evaluate MQG4AI interaction scenarios in part III.

5.2. Use Cases

This section outlines two very different medical and AI scenarios that comprise the foundation to evaluate MQG4AI in the next part III. The resulting MQG4DK and MQG4A visualizations can be consulted on GitHub[8] for a comprehensive illustration of MQG4AI.

First, a multi-label classification (fictional) ECG use case, situated in emergency medicine is contextualized. Analysis centered around a reliable performance metrics evaluation strategy [EM24] comprises the foundation for a complex leaf-QG compilation contribution to MQG4DK in chapter 7. The second use case is centered around a rare disease of the esophagus, Achalasia, and embedded within the medical software *EsophagusVisualization* for clinical research, as well as support of diagnosis and treatment [EM23]. Concretely, to illustrate MQG4A in chapter 8, we focus on model selection of segmenting the esophagus' shape in timed-barium esophagogram images (TBE). This comprises an elementary step within the software's workflow, highlighting the human-in-the-loop approach. For both use cases, we aim to extract a reliable design decision-making workflow for different components of the AI lifecycle during model development.

[8] `https://github.com/miriamelia/MQG4AI/blob/main/README.md`

5.2.1. Electrocardiogram Multi-label Classification – Fictional Use Case *Custodian* in Emergency Medicine

Cardiovascular diseases (CVD) are the primary cause of death worldwide and disproportionately affect patients with other health conditions, like diabetes or chronic kidney disease. Addressing the prevention, early detection, and treatment of CVD is a major challenge in healthcare, including emergency medicine (EM). [MS23, 2] EM comprises a *medical specialty* [ISO21b], and related AI application scenarios include continuous monitoring systems, intelligent enhancement of complication or outcome predictions, as well as triage systems [ISO21b, 6], with the latter providing structured processes to prioritize patients based on the severity of their condition and the urgency of treatment required. For instance, in emergency settings, AI-driven models can predict, during the triage process, whether a patient will need an ECG. [MS23, 2] Complementing time efficiency and resource allocation optimization, AI can assist in challenging diagnosis that take place in EM. For instance, myocardial infarctions are particularly difficult to manage, with an estimated 10.000 to 50.000 cases missed each year in US emergency departments. Generally, EM care costs are rising steadily in developed nations. In these fast-paced, high-pressure environments, emergency medical personnel must make quick decisions with limited information, leading to a high likelihood of diagnostic errors. This creates a clear need for (intelligent) CDSS tools in EM. [GS22, 1]

This section outlines the medical setting of our fictional use case, highlighting EM, the functioning of the human heart, CVDs, and ECGs, as well as their challenging, yet crucial interpretation. In addition, we shed light on the role of AI with respect to ECGs. Namely, we discuss AI application scenarios centered around ECGs before outlining challenges of intelligent enhancements for ECG analysis. Finally, we introduce our fictional *Custodian* use case, that is centered around an application-specific label structure.

5.2.1.1. Medical Background

EM became a medical specialty mainly due to "[...] the presence of patients with increased mobility who required unscheduled care that the current system could not accommodate (and increased financial support for these visits)" [HM22, 418]. The field is characterized by a high growth in number of emergency physicians [HM22, 418], as well as a broad combination of subspecialties (e.g. Medical Toxicology, Anesthesiology Critical Care Medicine, Internal Medicine-Critical Care Medicine, or Pain Medicine, among other disciplines). Therefore, EM is continuously aligned with medical and technological progress [HM22, 419], including CVDs and knowledge on the human heart.

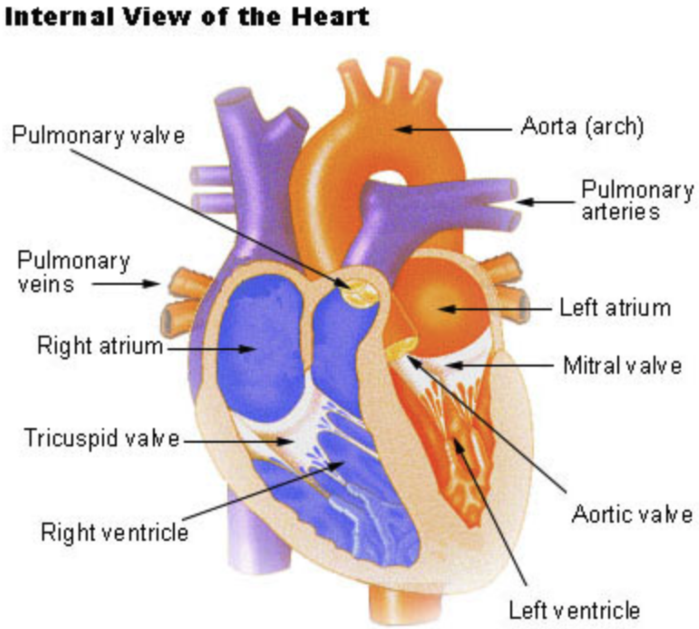


Figure 5.2.: The anatomy of the human heart, as depicted by [Nata].

The Heart Despite existing medical knowledge, there are still unresolved questions, and "[a]dvanced insights into disease mechanisms and therapeutic strategies require a deeper understanding of the molecular processes involved in the healthy heart" [LM20, 466]. The human cardio (heart) is a critical and complex organ, "[...] composed of four morphologically and functionally distinct chambers [...]" [LM20, 466]. The heart's anatomy is depicted in Figure 5.2. Generally, the four chambers comprise two thin-walled atria, that receive blood from veins, and two thick-walled ventricles, that pump blood forcefully. The right atrium receives deoxygenated blood from systemic veins, while the left atrium receives oxygenated blood from the lungs. Valves ensure one-way blood flow, and the two types comprise atrioventricular (tricuspid and mitral) valves between atria and ventricles and semilunar (pulmonary and aortic) valves at the bases of large vessels exiting the ventricles. Blood flows through the heart in a coordinated manner, with both sides working simultaneously. The right side pumps deoxygenated blood to the lungs, while the left side pumps oxygenated blood to the body. The myocardium, the heart's muscular wall, relies on an extensive blood and nutrients supply via the coronary arteries to maintain its function. Therefore, cardiac muscle is supported by an extensive network of blood vessels (arteries and veins) that deliver oxygen to the contracting cells and carry away waste products. [Nata]

Cardiovascular Diseases As stated by the World Health Organization (WHO), CVDs, i.e. "[...] a group of disorders of the heart and blood vessels" [Wor], are the leading global cause of death, accounting for 17.9 million fatalities in 2019, 85%

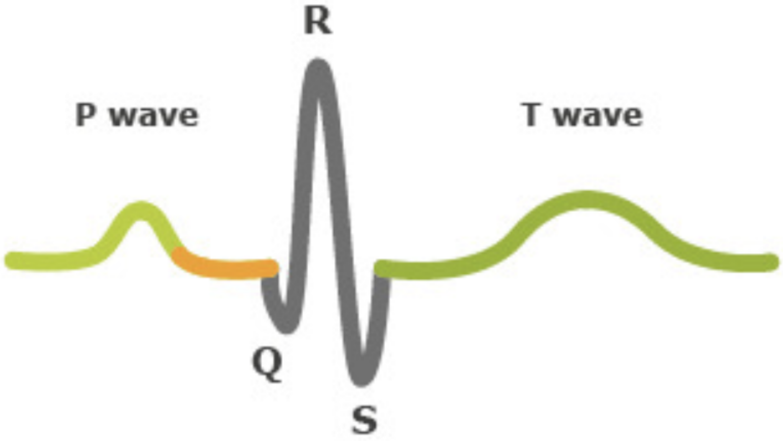


Figure 5.3.: Normal ECG pattern of a healthy heart, as depicted by [Ins23].

of which were from heart attacks and strokes. [Wor] The classification of CVDs depends on various factors, including their localization in the body, the presence of blockages or disrupted blood flow, and underlying causes such as structural abnormalities or electrical dysfunction, as well as the type of tissue affected, for instance. Ischaemic heart disease (heart attacks) and cerebrovascular diseases (strokes) are equally the "[...] leading cause of death in the EU" [eurb] according to Eurostat. The latter occurs, when arteries in the brain become clogged, and blood supply of the brain is limited. The thus resulting lack of oxygen causes brain cells to die, which can have severe consequences. Coronary artery disease (which typically results in cardiac ischaemia), manifesting acutely as myocardial infarctions, is caused by blocked blood flow in arteries supplying oxygen and nutrients to the heart. This blockage can cause damage or destruction to parts of the heart muscle, as insufficient blood flow leads to tissue death, and consequently weakens the heart. Both conditions require immediate medical attention. Most CVDs are preventable by addressing risk factors like tobacco use, unhealthy diets, obesity, physical inactivity, alcohol misuse, and air pollution. Early detection is vital to initiate effective management through counseling and medication. [Wor]

Electrocardiogram Comprising an "[...] important diagnostic technique [...]" [GA24a, 1634], ECGs play a crucial role in assessing heart health, and provide valuable insights into its condition, as well as help identifying potential indicators of various CVDs. They measure the heart's electrical activity using electrodes placed on the skin, and "[...] provide real-time visual representations of the heart's electrical movement throughout its cycle" [GA24a, 1635]. Figure 5.3 illustrates the ECG pattern generated by a steady heart beat. The P wave represents the electrical signal spreading through both atria, prompting them to contract and pump blood into the ventricles before relaxing. The QRS complex reflects the electrical impulse reaching the ventricles, causing them to contract. Finally, the T wave indicates the end of the electrical activity as the ventricles relax. [Ins23]

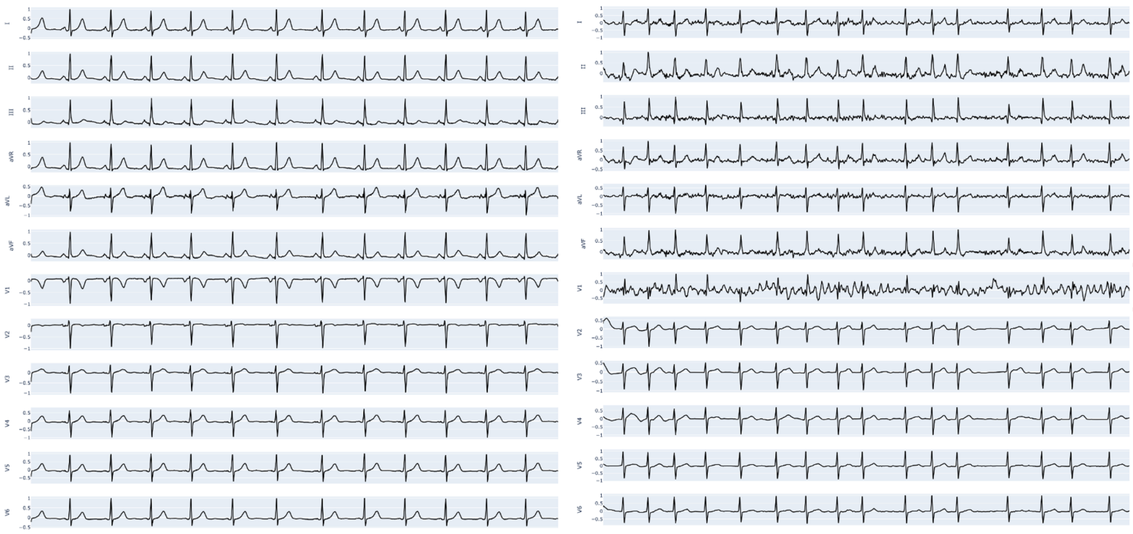


Figure 5.4.: Two 12-lead ECG recordings representing a regular heart beat (sinus rhythm) on the left, and an irregular heart beat (atrial fibrillation) on the right.

The standard 12-lead ECG uses ten electrodes, with six placed on the chest and one on each forearm and calf to "[...] record cardiac activity from multiple perspectives" [GA24a, 1635]. They are connected to an ECG-machine, that records the ECG graph for analysis, as depicted in Figure 5.4 for sinus rhythm and atrial fibrillation. The three types of ECG tests include resting ECGs while lying down, or exercise ECGs that are measured during physical activity, for a few minutes respectively. Holter monitor tests are based on a portable device that records the heart's activity over 24 hours or longer. [Ins23]

ECG Interpretation Diagnosing heart-related conditions can be complex, often requiring extensive manual examinations. To overcome this challenge, fast and accurate diagnostic tools are essential. While advanced techniques like Magnetic Resonance Imaging (MRI), Coronary CT Angiograms, Blood Tests, Coronary Angiography (CAG), and Echocardiograms are available, ECGs remain a simple and dependable method for diagnosing heart issues. By analyzing a 12-lead ECG, healthcare providers can identify a range of cardiac conditions, including arrhythmias, ischemia (reduced blood flow), and blockages, based on the provision of vital insights into the heart's rhythm, rate, and overall function. [GA24a, 1635] Challengingly, some of all existing conditions, summarized as heart arrhythmias/rhythms and other heart diseases (HARHD) [EM24, 2] are independent of one another and can coexist. Also, from a medical perspective, an ECG diagnosis alone often does not determine whether therapy is needed or what type (medication or electrical) is appropriate. Instead, ECGs contribute information to the severity of the diagnosis. For instance, ventricular tachycardia can cause heart rates ranging from 120 to 240 beats per minute,[9] and the patient's clinical condition can range from being asymptomatic to unconscious. Consequently, diagnoses and subsequent therapies frequently rely on a combination of multiple factors, including (combinations of) in an ECG identified HARHD. [EM25c]

Overall, ECGs are an indispensable tool that provides critical insights, especially in emergency medical situations, directly influencing patient management decisions. Due to its importance, healthcare professionals universally recognize ECG interpretation as an essential skill and include it as a key component of medical education. However, the medical community has voiced increasing concerns about potential shortcomings in ECG interpretation among healthcare providers. [KA23, 2] This necessitates comprehensive and continuous ECG training among healthcare professionals, especially in low- and middle-income countries, where three-quarters of CVD-related deaths occur [Wor]. For instance, during a study conducted in Ethiopia, three-fourths of pediatric and child health residents reported inadequate ECG training during their residency, leading to a lack of confidence in interpreting ECGs. This resulted in missed ECG diagnoses of common life-threatening conditions. [MH24, 1] Finally, the growing use of AI provides a means to support and enhance the interpretation of ECGs worldwide, while it is not meant to replace the expertise, knowledge, and experience of trained medical professionals [KA23, 14].

5.2.1.2. Cardiovascular Conditions, ECGs & AI

While currently, AI for ECG interpretation and classification displays potential, there is insufficient evidence to indicate that it outperforms existing ECG interpretation software [KA23, 14]. Therefore, the focus of ECG-AI currently lies on the lifecycle development and data stages, in contrast to more translational deployment and maintenance perspectives. Also, existing intelligent ECG-systems "[...] tend to prioritize outcomes outside the scope of comprehensive 12-lead ECG interpretation" [KA23, 14]. In addition, effective ECG analysis in medical practice relies on factors such as patient history, physical examination, and clinical judgment, which current models do not fully capture. [KA23, 14] Overall, as AI technology advances and more data becomes available, its role in ECG diagnosis and management is expected to grow, helping clinicians provide better care [MS23, 1]. Therefore, as AI becomes more integrated into clinical practice, training programs should ensure medical professionals are equipped with the skills to effectively use this technology, while also understanding its benefits and limitations [KA23, 14].

Application Scenarios Various scenarios how to apply AI surrounding ECGs are possible, and "[...] AI may have many use cases to improve many current clinical

[9] A normal resting heart rate for adults usually lies between 60 to 100 beats per minute.

processes where an ECG is involved to deal with cardiovascular diseases" [MS23, 10]. AI capabilities reach from interpretation and detection of ECG abnormalities, over ECG signal processing, aiming to improve quality and accuracy, to risk prediction and therapy guidance, possibly including further clinical variables in addition to ECGs as model input [MS23, 2]. Overall, the concrete application scenario depends on a multitude of different factors concerning decisions on the model's desired capabilities, medical domain knowledge tailored to the desired model output, as well as the clinical target workflow of the intended use. For instance, since there is "[...] a lack of high-performing models for the diagnosis of myocardial infarction in real world scenarios" [GS22, 1], in [GS22], the authors propose a "deep neural network based on convolutional layers similar to a residual network" [GS22, 1], and include additional clinical variables (i.e. age and sex) as model input tailored to one particular heart disease and intended for utilization in emergency departments. Therefore, the model output comprises three distinct and mutually exclusive labels, including "[...] non-ST-elevation myocardial infarction (NSTEMI), ST-elevation myocardial infarction (STEMI), and control status [...]" [GS22, 1]. Another example is demonstrated in [ZA25], where the authors test ChatGPT's performance regarding ECG interpretation in comparison to physicians. They found that it displays moderate accuracy, and discrepancies, particularly in evaluating critical cases, restrict its applicability, while advancements in research and technology could improve its reliability, making it a valuable support tool for emergency physicians in the future [ZA25, 1]. Finally, our fictional use case is located at the interface of ECG interpretation and therapy recommendations. Its *Custodian* functionality is shaped by the process step within the concrete EM setting, aiming to classify patients at risk, and to provide guidance on reasonable medical follow-up steps to support the respective healthcare professional in charge. Among other design decisions, this particular setup influences the choice of predicted labels, or combination of HARHD, as further outlined in chapter 7.

Challenges The previously introduced general medical AI challenges equally apply to the ECG domain, and we highlight a combination of more specific technical layers surrounding data quality and distribution, fitting model selection among the multitude of existing algorithms, as well as ECG-AI explainability, which are not focus of our MQG4AI lifecycle blueprint evaluation in chapter 7. There, we outline the need for accurate performance measurement in more detail, as theoretical foundation of our proposed approach to MQG4DK contributions. Reliable performance evaluation metrics are strongly influenced by the data distribution and use case, as well as comprise the foundation for further (design) decision-making along the AI lifecycle.

1. *Data* Due to the previously introduced intricacies surrounding medical ECG interpretation, data set creation is not a trivial, but necessary task. As a result, the quality of annotations needs to be assessed carefully, and inter-annotator agreements can support this process [BN20]. Addressing the

costliness of time-consuming ECG labeling, for instance, architectures exist that include active learning (AL), where the model defines which samples it needs labeled to reduce the size of annotated training data [SF23]. Depending on the application scenario, which may include binary, multi-class (predict one of three or more possible labels), or multi-label (predict multiple labels that may coexist) scenarios, the precise labeling structure needs to be considered from the start of data set creation, while data sets may function as foundation for multiple ECG-AI scenarios. One ECG recording, or data sample "[...] can include multiple, differently correlated heart arrhythmias/rhythms and other heart diseases [HARHD], or labels from a technical viewpoint" [EM24, 2], which is introduced for our fictional use case in the next section. Further, ECG labels are characterized by high data imbalance, with healthy patients, represented by e.g. a sinus rhythm and the absence of other HARHD being more prevalent within the data. This tendency, equally found in open-source data sets, such as the related 2020 and 2021 PhysioNet challenge datasets,[10] impacts model performance and needs to be considered during evaluation metrics selection. [EM25c]

2. *Model Selection* In addition to suitable data set creation, finding the right model architecture for the use case at hand is required. For instance, the authors of [GA24a], provide a comprehensive overview on "[...] recent advancements in [heart disease] classification using various machine learning, deep learning, and ensemble learning algorithms [...]" [GA24a, 1634], focusing on "[...] various methods' strengths, limitations, and advancements [...]" [GA24a, 1634]. DL-based models for ECG-AI are mainly based on CNNs, e.g. based on ResNet, "[...] a popular image classification architecture" [BA20, 883] [GS22, 4], and may additionally consider RNN, and LSTM architectures [GA24a, 1639].

3. *Explainability* The increasing complexity of DNN architectures may result in prioritizing accuracy over explainability. This shift has raised concerns among clinicians, as a lack of focus on explainability may hinder the clinical adoption of advanced AI-ECG tools. [AZB23, 292] Therefore, as further outlined in [AZB23], XAI is a crucial component towards risk mitigation of intelligent (ECG) systems, supporting the reliability of lifecycle design decisions, bias mitigation, and "[...] AI-ECG interpretations need to be explained to allow the discovery of new patterns and mechanistic links to disease pathophysiology" [AZB23, 292], among other factors. For instance, regarding the previously broached myocardial infarction multi-class classification model, the authors applied Grad-CAM plots to shed light on the model's inner workings [GS22, 5].

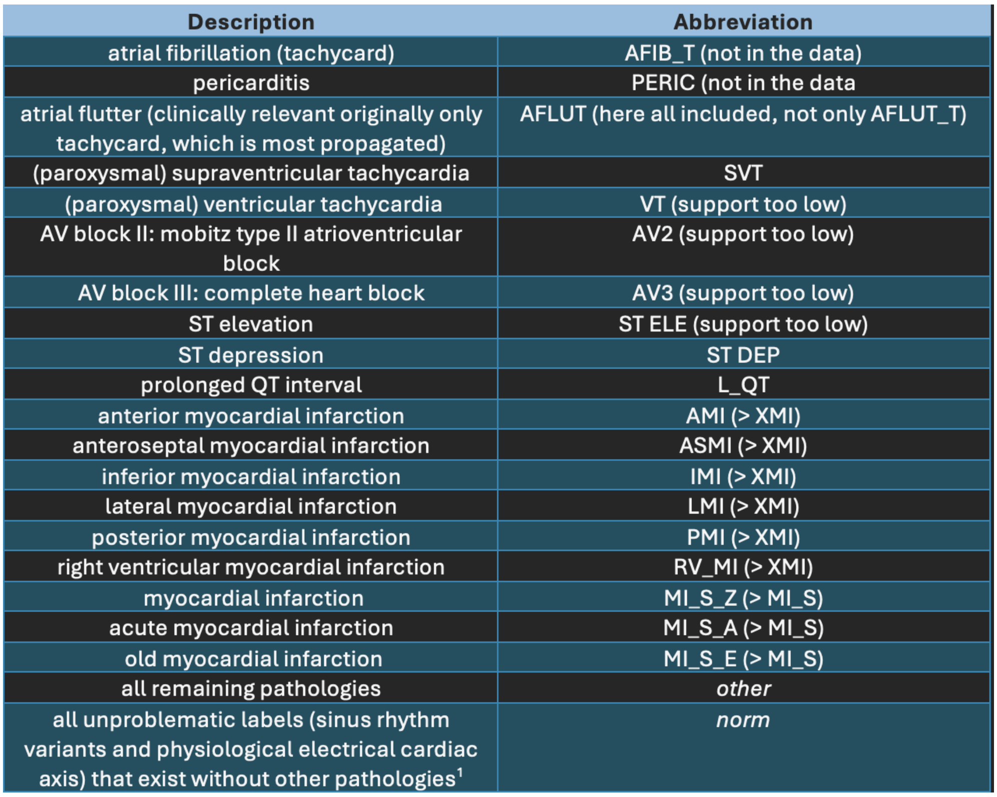
[10] The 2021 data set builds on the 2020 data set and consists of a combination of different, partly undisclosed sources [RM21a, 2]. It is the "[...] largest freely available repository of standard 12-lead ECG records and consistent annotations for 30 clinical diagnoses of cardiac abnormalities" [MS23, 6].

Finally, in chapter 7, we outline how to approach metrics selection tailored to the previously described multi-label ECG classification fictional use case *Custodian* in more detail, aiming to extract generalizable process steps and design decisions. We address the risk unreliable performance evaluation metrics, and illustrate a MQG4DK contribution scenario, which is based on a leaf-QG compilation, aiming for reliability by design. Therefore, the next section introduces our use case-specific label structure, which is based on Physionet data and links important pre-selection information for evaluation metrics.

5.2.1.3. Fictional *Custodian* ECG Multi-label Structure

The fictional use case that comprises the foundation for the evaluation of reliable multi-label performance evaluation metrics in chapter 7 is "[...] an alarming guard functionality anywhere patients at risk could appear [...]" [EM24, 4]. "Approximately 20% of emergency department (ED) visits involve cardiovascular symptoms" [ZA25, 1], therefore, as previously introduced, we highlight EM, as intended environment of use. Additional, possible scenarios include ambulances and general practitioners' offices, or other settings, where medical personnel is not specialized in ECG interpretation. The *Custodian* AI's primary goal is to enhance patient welfare by enabling early identification of cardiological issues, as timely intervention is often crucial. The envisioned intelligent medical product uses multi-label classification of 12-lead ECGs to assist in diagnosing patients. Therefore, one ECG sample may contain multiple, co-existing HARHD (or, data labels) that may correlate with one another. Through collaboration between the medical team and the intelligent system, critical diagnoses are expected to improve, leading to better outcomes for patients in form of clear guidelines for action embedded within the clinical emergency workflow. For instance, follow-up steps after classification may result in immediate medication, or referral to a hospital for further diagnostics, depending on the model's prediction and physician's evaluation. This setting directly shapes the label structure that the model predicts, as described in the following.

Multi-label Structure The proposed label structure for multi-label classification of ECGs integrates an alarming guard functionality and is built upon the open-source Physionet dataset [RM21a], where ECG labels correspond to correlated HARHD. For use case-adaptation, the proposed label structure is designed to group similar labels regarding the clinical follow-up step that is required directly after a label is predicted. It is described in detail in the supplementary material of our corresponding publication [EM25c] [EM24]. Overall, this structure ensures that labels guide immediate preclinical decisions, particularly in emergency settings, where rapid interventions (e.g. medication or electrical therapy) are critical. From a medical perspective, some labels are independent and can coexist, while others indicate distinct diagnostic pathways. Importantly, the presence of one la-



Description	Abbreviation
atrial fibrillation (tachycard)	AFIB_T (not in the data)
pericarditis	PERIC (not in the data
atrial flutter (clinically relevant originally only tachycard, which is most propagated)	AFLUT (here all included, not only AFLUT_T)
(paroxysmal) supraventricular tachycardia	SVT
(paroxysmal) ventricular tachycardia	VT (support too low)
AV block II: mobitz type II atrioventricular block	AV2 (support too low)
AV block III: complete heart block	AV3 (support too low)
ST elevation	ST ELE (support too low)
ST depression	ST DEP
prolonged QT interval	L_QT
anterior myocardial infarction	AMI (> XMI)
anteroseptal myocardial infarction	ASMI (> XMI)
inferior myocardial infarction	IMI (> XMI)
lateral myocardial infarction	LMI (> XMI)
posterior myocardial infarction	PMI (> XMI)
right ventricular myocardial infarction	RV_MI (> XMI)
myocardial infarction	MI_S_Z (> MI_S)
acute myocardial infarction	MI_S_A (> MI_S)
old myocardial infarction	MI_S_E (> MI_S)
all remaining pathologies	*other*
all unproblematic labels (sinus rhythm variants and physiological electrical cardiac axis) that exist without other pathologies[1]	*norm*

Figure 5.5.: The proposed label structure, as incorporated by *Custodian*, mapped to Physionet data, and reflecting 8 and 20 distinct labels based on support in the data.

bel does not necessarily imply the absence of other conditions. Since the dataset lacks a label explicitly indicating the absence of ischemic signs, two *Container Labels* were introduced:

1. *Norm* Includes all labels that indicate a healthy patient, and in the absence of HARHD, it is assigned to the respective ECG sample. For example, a patient may have a sinus rhythm alongside ST elevation, which could indicate an acute cardiac infarction. Therefore, the *Norm* label is only assigned when no pathology is present.

2. *Other* Comprises unspecific diagnoses that do not directly lead to immediate intervention but require further clinical investigation. In ECG diagnostics, conditions such as ventricular tachycardia do not necessarily dictate whether treatment is required or which type (medication vs. electrical therapy) is appropriate. Therefore, the adapted label structure primarily focuses on labels triggering immediate preclinical interventions, such as arrhythmias and ischemic signs. All other cardiac pathologies are grouped under *Other* and require additional assessment. Consequently, if this label appears, further investigation is necessary.

The resulting label structure defines possible label combinations, and *Other* and pathologies can coexist, meaning additional conditions may be present. *Norm* is predicted only when no other labels are present, ensuring that it strictly represents a healthy ECG. To align the data with our clinical needs, the use case-adapted label structure was mapped to the Physionet dataset using *Systematized NOmenclature of MEDicine – Clinical Terms* (SNOMED CT)[11] codes. However, this process revealed incompleteness, as some clinically relevant labels do not exist in the dataset. Figure 5.5 depicts mapping the medical label structure to the Physionet data, and summarizes label abbreviations for 20 (and 8) labels based on their dataset support. XMI and MI_S are grouped *Container Labels*, as they correspond to similar clinical follow-up steps. In EM, precise infarction localization is not always relevant, making this grouping clinically appropriate.[12]

In summary, key considerations in label definition are based on adapting the classifier's desired output to the specific *Custodian* use case. Labels are organized through domain translation, ensuring alignment with the fictional real world clinical setting and interpretation requirements. This structured approach ensures that the model provides clinically meaningful predictions, supporting both immediate intervention and further diagnostic evaluation when needed. After introducing our fictional *Custodian* use case, centered around ECGs and heart conditions in EM, which addresses an omnipresent, global concern, the next section outlines a very different medical scenario: the rare disease Achalasia of the esophagus and the medical software *EsophagusVisualization*.

5.2.2. Intelligent Enhancement of *EsophagusVisualization* – A Medical Software towards Trustworthy Precision Medicine for the Rare Disease Achalasia

Achalasia represents a rare chronic motility disorder of the esophagus, characterized by absence of deglutitive lower esophageal sphincter (LES) relaxation and impaired propulsive tubular peristalsis [BG14], as depicted in Figure 5.6. This results in difficulties regarding ingestion, as well as discomfort for the patient. Aiming to support medical research, diagnosis and treatment of Achalasia, our experimental software is centered around a novel data type in form of a dynamic and interactive 3D-reconstruction of the esophagus based on multi-modal medical input data. A special challenge of this use case comprises its categorization as a rare disease, which results in tendencially low support. Meaning, digital, yet intelligent enhancement does not exist tailored to the disease, which is the case

[11] `https://www.snomed.org/what-is-snomed-ct`

[12] With respect to the metrics selection process, as proposed in chapter 7, due to insufficient data support, per-label metrics could not be calculated for PERIC, AFIB_T, several fine-grained ischemia labels, AV2, AV3, ST ELE, and VT. Therefore, these labels were excluded from our evaluation example, while they are still relevant for the fictional use case from a medical view.

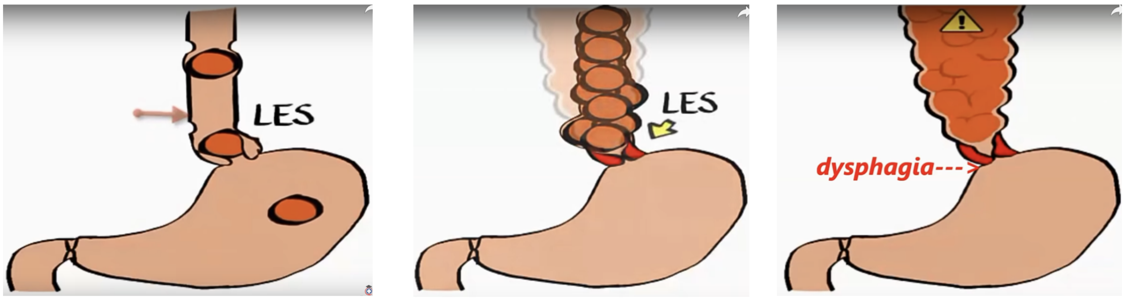


Figure 5.6.: Concerning a healthy person, food passes the lower esophageal sphincter (LES) to enter the digestive tract. This mechanism is interrupted for Achalasia patients, resulting in dysphagia, regurgitation and weight loss, illustration copied from [Cha17].

for Achalasia. The software is developed in the context of research and higher education at the University of Augsburg. Our project comprises multiple long-term visions that result in a holistic approach tailored to the rare disease Achalasia.[13]

Finally, the tool is already utilized for medical research by the Department of Gastroenterology of the University Hospital Augsburg (UHA) [GV24]. We aim to quantify treatment success based on software-specific calculated indexes that capture the interplay between pressure and shape along the esophagus, highlighting differences in behavior between the LES and tubular esophagus (TE), as illustrated in Figure 5.7. In addition, the interdisciplinary project is developed in cooperation with the Chair of Moral Theology at the University of Augsburg and the Center for Responsible AI Technologies (CReAITech) [JS23]. This section introduces *EsophagusVisualization*, the software's setup, vision, and limitations, highlighting the role of AI.

5.2.2.1. Medical Background

Achalasia is a rare esophageal motility disorder with an incidence of approximately 1.6 to 2.5 cases per 100.000 individuals per year. It affects men and women equally, with a 1:1 ratio. The disease exhibits a bimodal age distribution, with two distinct peaks in incidence: the first occurring during the 3rd to 4th decade of life, and the second after the age of 60. Achalasia is characterized by a progressive disease process that worsens over time if untreated. The exact cause of Achalasia remains uncertain, but several potential mechanisms are under investigation. They include autoimmune factors in triggering esophageal dysfunction, infectious agents that contribute to the disease, hereditary factors based on

[13] The project is compliant with the Declaration of Helsinki. The study protocol was approved by the medical ethics committee of the Ludwig Maximilian's University of Munich on June 27, 2022 (registration no. 22-0149).

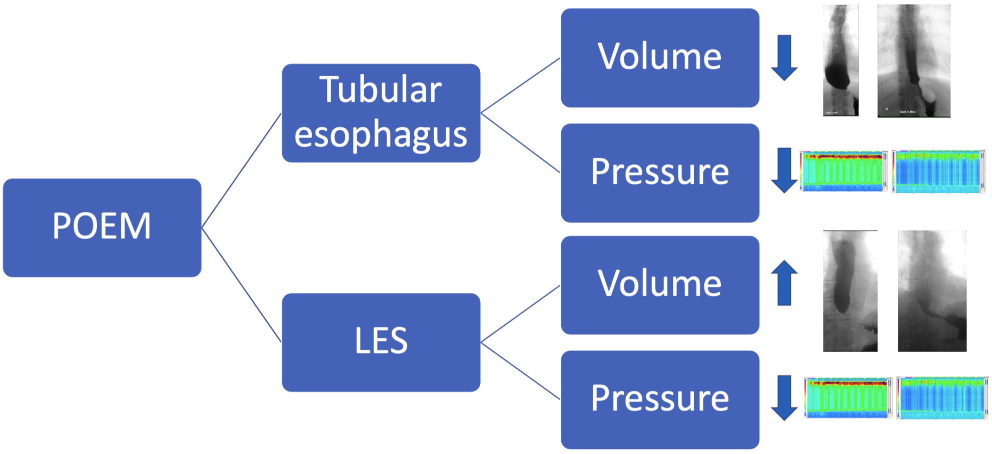


Figure 5.7.: Medical Research with the *EsophagusVisualization* tool, aiming to capture and analyze the esophagus' behavior post-therapy (POEM), as depicted and introduced in [GV24].

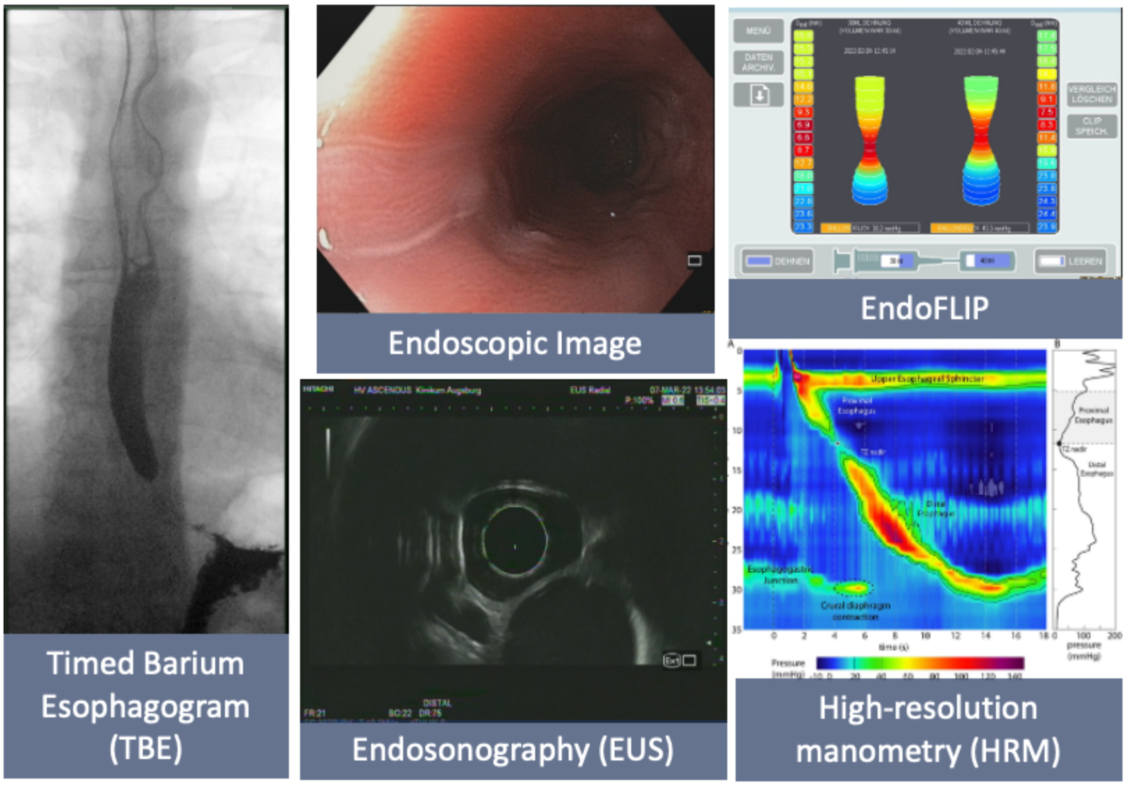

Timed Barium Esophagogram (TBE)
Endoscopic Image
EndoFLIP
Endosonography (EUS)
High-resolution manometry (HRM)

Figure 5.8.: Clinical raw data surrounding Achalasia (source: UHA).

genetic predispositions or familial clustering, and degenerative processes of the esophageal nervous system. Particularly challenging for diagnosis and treatment is the existence of different Achalasia types and treatment options, in addition to the current state of medical knowledge on the disease.

Diagnosis & Monitoring The diagnosis of Achalasia relies on a combination of clinical evaluations and advanced diagnostic tools, as depicted in Figure 5.8:

- *Esophagogastroduodenoscopy (EGD)* to rule out structural abnormalities and assess esophageal function in form of endoscopic images (or videos).
- *Timed Barium Esophagogram (TBE)*, a radiological swallowing test to visualize esophageal emptying and dilation, accessible as image data.
- *Esophageal High-resolution Manometry (HRM)*, which is the gold standard for measuring esophageal pressures and confirming impaired motility as comma-separated values (.csv).
- *Endolumenal Functional Lumen Imaging Probe (EndoFLIP)*, a functional lumen imaging probe used to evaluate esophageal compliance and distensibility, resulting in .csv input data.
- *Endoscopic Ultrasound (EUS)* to assess esophageal wall thickness and structural abnormalities, in form of pictures (or videos). Currently, this data type is not considered for 3D-reconstruction, but would provide valuable information on esophageal cross-sections, as well as wall thickness.

Treatment Options & Success The response to treatment is influenced by factors such as type of Achalasia, degree of dilatation, kinking of the TE, pressure at the LES, and its distensibility. Achalasia types and their prevalence are depicted in Figure 5.9. Adding to this complexity, multiple treatment options exist, with diverging success depending on the individual patient's constituency. Generally, treatment is centered around relieving pressure in the LES to regain swallowing capabilities. Peroral endoscopic myotomy (POEM) is currently one of the standard treatment options, and steps 1 – 4, as depicted in Figure 5.10, are described as follows: During the endoscopic procedure, the endoscopist dissects the circular musculature via a submucosal tunnel as access, that comprise the border to sensitive organs such as the cardio (heart) and pulmo (lung). Overall, this procedure creates a "path" based on electrical burns, which runs from variable heights along the TE to the LES. The myotomy of the LES and parts of the tubular esophagus allows for relaxation at the LES. Finally, the mucosal entry to the submucosal tunnel is then closed with clips. Building on diagnostic tools, several studies have investigated various measures, including HRM, TBE, and impedance planimetry (EndoFlip), to predict clinical responses post-therapy and showed mostly moderate predictive power. For instance, regarding POEM, fine-grained information, such as the myotomy position at the anterior or posterior wall of the esophagus may impact treatment success.

As up to 20% of Achalasia patients experience symptom recurrence over time [MA16], there is a need to shift our therapeutic focus from solely evaluating clin-

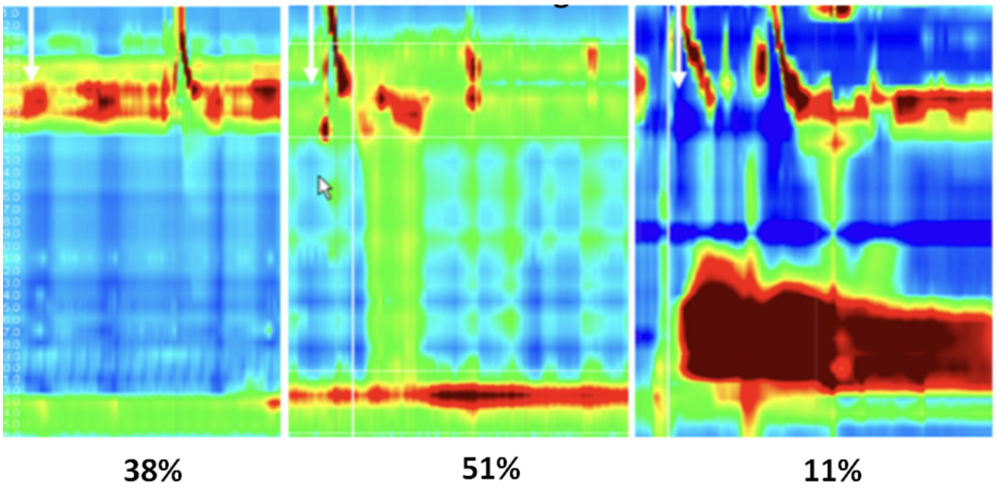


Figure 5.9.: Three different types of Achalasia according to the Chicago Classification of esophageal motility disorders [KP15], using high-resolution manometry (HRM). Here depicted including their probability of occurrence (source: UHA): type 1 aperistaltic (left), type 2 panesophageal pressurization (middle), type 3 spastic (right).

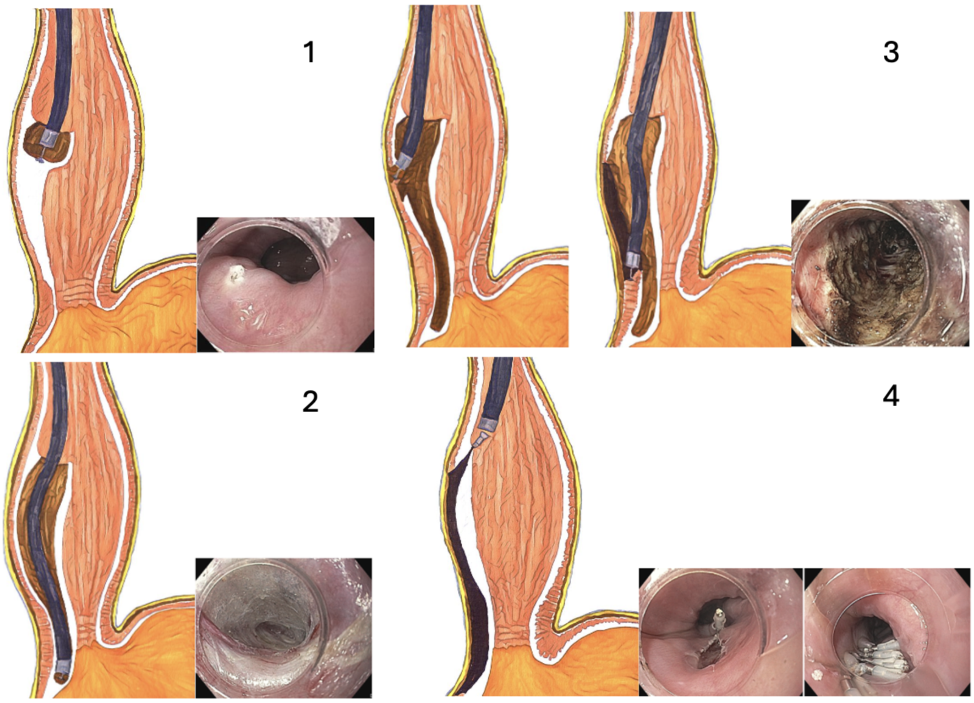


Figure 5.10.: Conceptual process steps of Peroral Endoscopic Myotomy (POEM), as depicted in [Dep]. POEM is a novel treatment option of Achalasia, and with promising early-study results [WY19].

ical responses to analyzing the underlying pathophysiological changes and remodeling processes induced by treatment. For instance, Figure 5.11 depicts pre- and post-therapy reconstructions. They are characterized by clear differences,

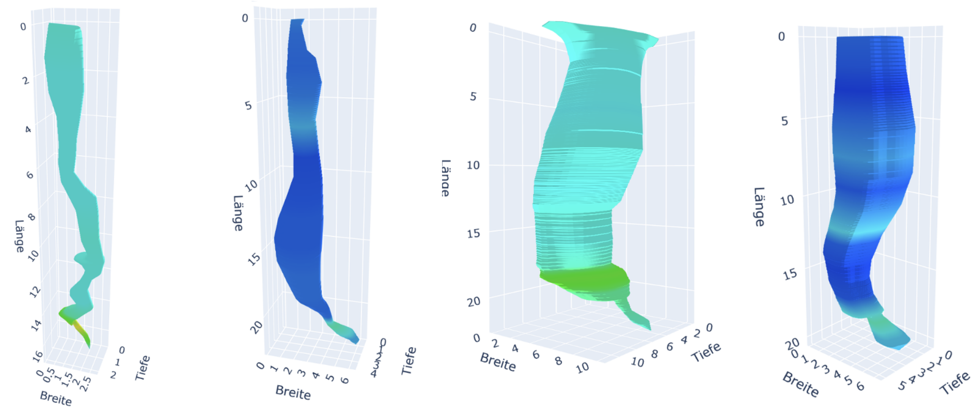


Figure 5.11.: Two possible therapy effects, depending on the respective patient's condition. The narrower initially, the broader post-therapy (left), and the broader initially, the narrower post-therapy (right) (source: UHA and *EsophagusVisualization* software).

and more broad analysis and medical research may lead to valuable observations. For instance, findings may indicate that the narrower the esophagus pre-therapy, the wider it is post-therapy, and vice versa. These dependencies may enhance a successful choice of treatment option tailored to the individual patient.

Overall, existing measures do not comprehensively capture the multifaceted changes in geometry, pressure, volume, and distensibility that occur following Achalasia therapy. This section summarizes the main motivations for the development of the *EsophagusVisualization* tool, including a complex and lengthy diagnostic process, and the need for medical research.

5.2.2.2. Novel Data Type – Multi-modal 3D-Reconstruction of the Esophagus

Currently, Achalasia patients iteratively undergo a number of different medical measurements, as illustrated in Figure 5.8, that are all recorded individually and the responsible physician single-handedly maps the results for a comprehensive image of the respective patient's situation. In addition, as previoudly introduced, a lot of information on Achalasia is still unknown from a medical viewpoint, resulting in the need for a multi-dimensional analysis that comprises all relevant factors centered around individual patients, their respective constituency, as well as medical data in relation to treatment success, supporting clinical research.

Aiming to enhance diagnosis, treatment, and medical research, since December 2021, the *EsophagusVisualization* tool, a first prototype, is being developed, resulting in a multi-modal 3D-reconstruction of the esophagus. The medical human-in-the-loop is integrated towards ensuring accuracy and transparency. The current

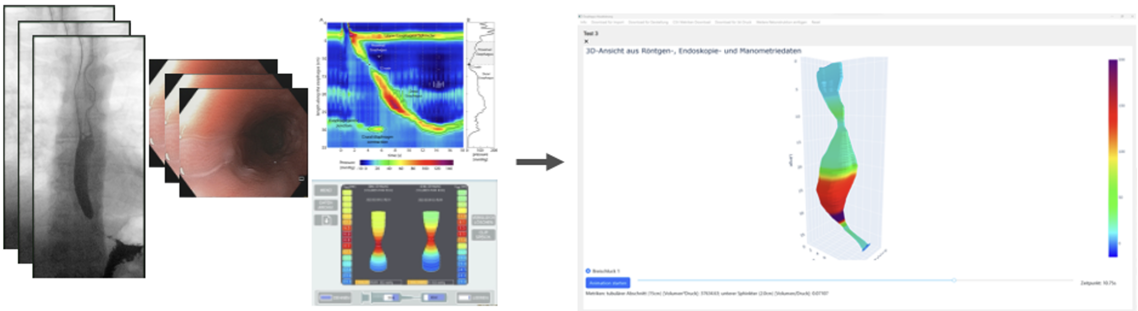


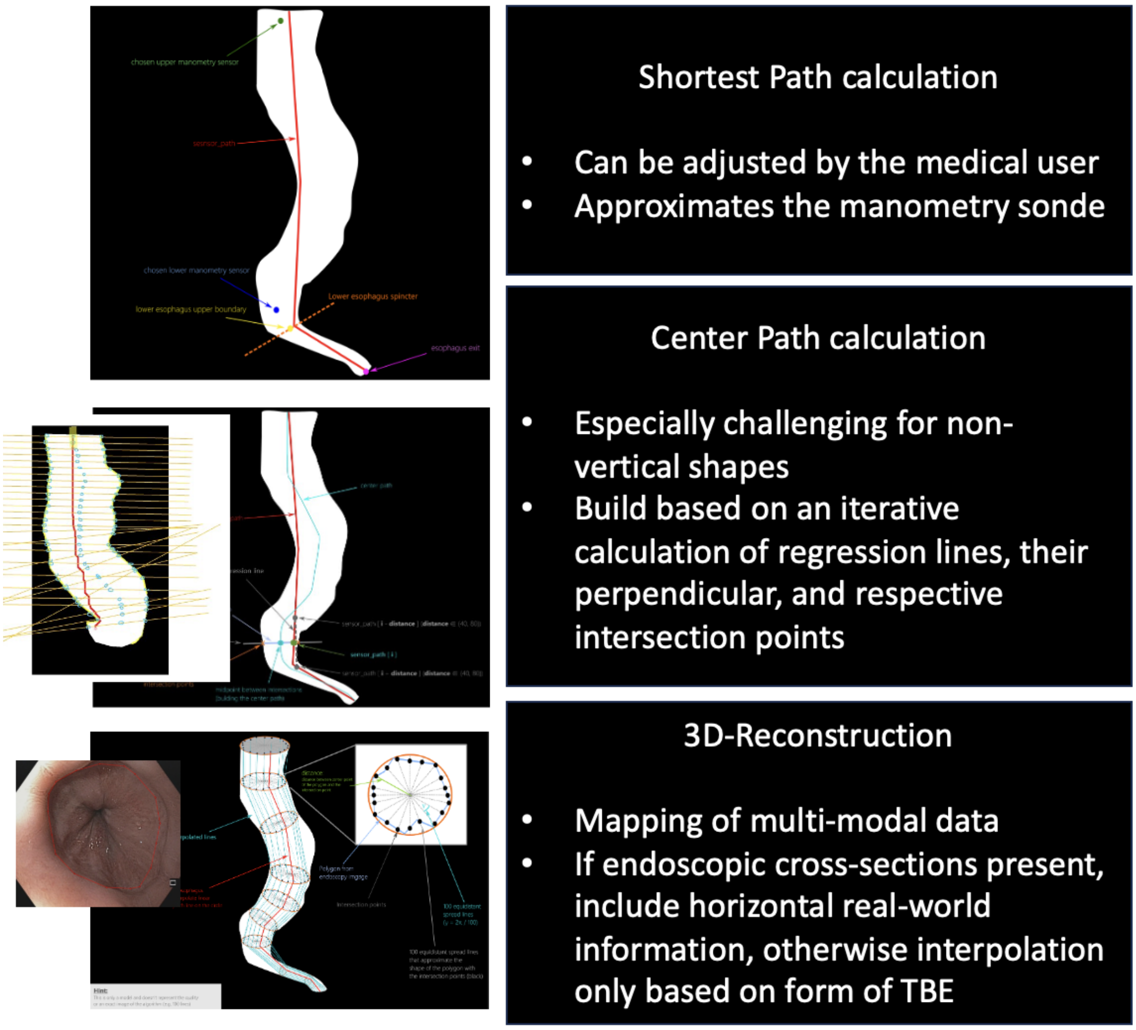
Figure 5.12.: Illustrated input of the *EsophagusVisualization* software for a multi-modal 3D-reconstruction of the esophagus (source: UHA and *EsophagusVisualization* software).

prototype (v3) supports four of the five relevant data types, parallel analysis of multiple reconstructions, as well as allows for the calculation of medical indexes that analyze pressure and volume [GV24], among other features, as illustrated in Figure 5.12.[14] Additionally, v3 provides the calculation of an *esophageal pressurization index*, which is based on the distal contractile integral (DCI) [KP15]. The DCI mirrors the peristaltic wave, which reflects the swallowing process, and there is no previously defined comparable value for Achalasia patients due to the disrupted tubular peristalsis. However, it is still reasonable to measure the disrupted tubular peristalsis alongside. It is calculated as the product of the mean contraction amplitude, the duration of the contraction, and the length.

3D-Reconstruction Algorithm At the heart of the tool lies the mapping algorithm that brings the combination of medical input information on shape and pressure of the esophagus together, resulting in an approximated 3D-reconstruction, as explained in Figure 5.13. Generally, the procedure is based on three main steps. First, the *shortest path* (or, *sensor path*) is calculated, aiming to represent the HRM tube for color mapping of pressure values in form of the HRM matrix onto the generated 3D model.[15] Next, the esophagus' *center path* is derived. Therefore, along the *shortest path*, for every point, regression lines are calculated with a variable distance of two points along the *shortest path*. This flexibility is necessary to

[14] Within the project, in close cooperation with our medical partners, and as foundation for higher eduction, additional features include different export functionalities, highlighting a .stl export for 3D-printing of the esophagus, which is, among other scenarios, applicable for medical education. In addition, great emphasis is placed on usability and the human-in-the-loop approach aiming for an accurate and trustworthy approximation, as will become clearer in the following sections.

[15] The color mapping of EndoFLIP data follows a similar approach, while necessitating some preprocessing. Extracting EndoFLIP data involves isolating key moments when the balloon is inflated to 30 ml or 40 ml from an XLSX file that records the entire examination. To approximate these critical points, sensor values at these inflations are aggregated, and statistical measures such as median, mean, minimum, and maximum are calculated to provide detailed insights. Additionally, the consideration of information about the distance of each sensor along the tube, which supports subsequent mapping is included.


chosen upper manometry sensor
sesnsor_path
chosen lower manometry sensor
Lower esophagus spincter
lower esophagus upper boundary
esophagus exit
center path
sensor_path [i - distance] |distance ∈ (40, 80)
sensor_path [i]
sensor_path [i - distance] |distance ∈ (40, 80)
intersection points
midpoint between intersections (building the center path)
distance
polated lines
Polygon from endoscopy image
Intersection points
100 equidistant spread lines
100 equidistant spread lines that approximate the shape of the polygon with the intersection points (black)
Hint:

Shortest Path calculation

- Can be adjusted by the medical user
- Approximates the manometry sonde

Center Path calculation

- Especially challenging for non-vertical shapes
- Build based on an iterative calculation of regression lines, their perpendicular, and respective intersection points

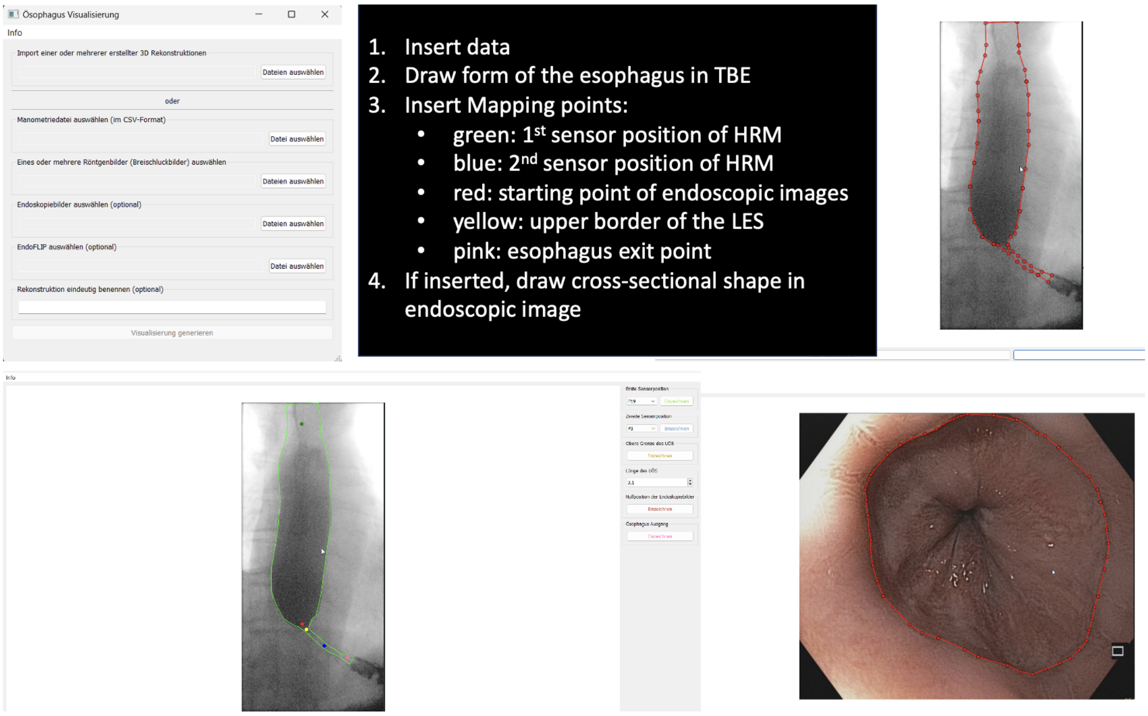
3D-Reconstruction

- Mapping of multi-modal data
- If endoscopic cross-sections present, include horizontal real-world information, otherwise interpolation only based on form of TBE

Figure 5.13.: The 3D-reconstruction algorithm that maps different input data together explained. In v3, the medical user can correct and adjust the calculated shortest (sensor) and center paths, assessing and enhancing reconstruction quality. (source: based on [GV24])

model varying degrees of skewed forms, which is common for Achalasia patients. Next, the perpendicular to the regression lines on the shortest path are calculated. Then the perpendiculars intersect the esophagus' boarder, and the middle points of the remaining lines are calculated, resulting in the *center path*. Finally, the 3D-reconstruction is generated through interpolation. If endoscopic images are present, the respective diameter of these cross-sectional views along different heights of the esophagus is included. Endoscopic input images therefore, are required to provide the height of their capture as their filename. Otherwise, the interpolation is based solely on information provided by the TBE image.

***EsophagusVisualization* Workflow** Aiming to enhance accuracy, as well as trustworthiness of the 3D-reconstruction, the medical human-in-the-loop plays a cru-


Ösophagus Visualisierung
Info
Import einer oder mehrerer erstellter 3D Rekonstruktionen
Dateien auswählen
oder
Manometriedatei auswählen (im CSV-Format)
Datei auswählen
Eines oder mehrere Röntgenbilder (Breischluckbilder) auswählen
Dateien auswählen
Endoskopiebilder auswählen (optional)
Dateien auswählen
EndoFLIP auswählen (optional)
Datei auswählen
Rekonstruktion eindeutig benennen (optional)
Visualisierung generieren
Info
Einzeichnen
Einzeichnen
Obere Grenze des UÖS
Einzeichnen
Nullposition der Endoskopiebilder
Einzeichnen
Ösophagus Ausgang
Einzeichnen

1. Insert data
2. Draw form of the esophagus in TBE
3. Insert Mapping points:
   - green: 1st sensor position of HRM
   - blue: 2nd sensor position of HRM
   - red: starting point of endoscopic images
   - yellow: upper border of the LES
   - pink: esophagus exit point
4. If inserted, draw cross-sectional shape in endoscopic image

Erste Sensorposition: P19

Zweite Sensorposition: P3

Länge des UÖS: 3,1

Figure 5.14.: 3D-reconstruction workflow, focusing on the medical human-in-the-loop within the *EsophagusVisualization* software. Note that the data input section has evolved in v3 aiming to capture more relevant patient data towards the creation of a comprehensive Achalasia data set, which is outlined in more detail in the next section, and depicted in Figure 5.16 (source: based on *EsophagusVisualization* software).

cial role within the reconstruction workflow. As depicted in Figure 5.14, first the patient data is inserted and the esophagus' shape marked in the TBE image, based on which all other data is mapped together. Next, mapping points are inserted, so that the algorithm "knows" how to combine all input information to generate the 3D-reconstruction. The blue and green points mark the HRM sensor positions at the beginning and the end to combine shape and pressure information. In addition, the red point identifies the starting point of endoscopic images, if existent. Therefore, the physician is required to take the picture chain bottom-up during the endoscopy, and to include their respective height as file name. In our scenarios, endoscopic images were taken with a distance of approximately 2 cm. Further, to enhance the algorithm's accuracy, the upper border of the LES is inserted (yellow), as well as the esophagus' exit point (pink). Finally, the esophagus diameter is marked in the endoscopic images, if present, before generating the 3D-reconstruction.

Human-in-the-Loop & Validation To ensure validity and robustness regarding individual 3D-reconstructions, the human-in-the-loop workflow is incorporated by design. This approach functions as quality-check of the generated 3D-reconstructions, through professional human oversight of intermediate steps that

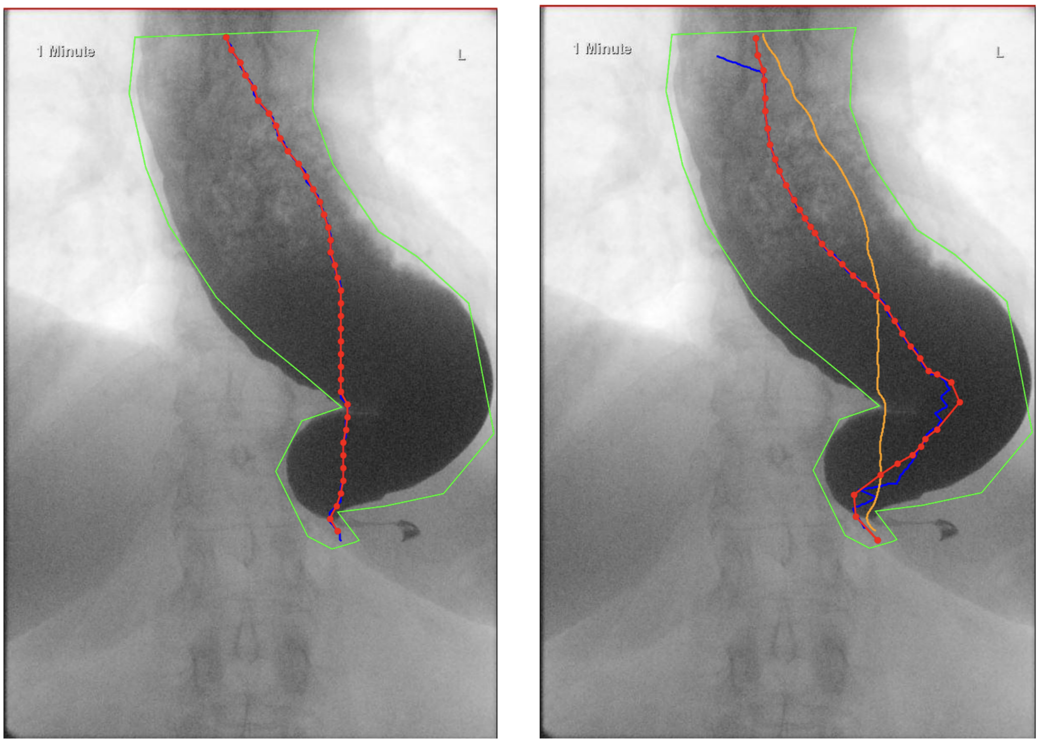


Figure 5.15.: The calculated sensor path, which can be manually adjusted (left). The thus derived center path (right, blue), and its manually adjusted version (right, red). (source: *EsophagusVisualization* software)

directly impact the final result. Complementing the previously introduced workflow, where the medical user manually inserts TBE and endoscopic shapes, as well as mapping points, the calculated sensor and center paths are visualized within v3, and can be adjusted by the medical user for a more accurate mapping, as visualized in Figure 5.15. Highly skewed shapes, which is common among Achalasia patients, may impact algorithm accuracy, and manually adjusting the sensor and center paths, is aimed at further enhancing and monitoring 3D-reconstruction data quality.

Approximation & Testing Generally, the resulting 3D-reconstruction is an approximation of the respective patient's esophagus, with the principal aim to support physicians in providing patient well-being. The accuracy of the thus generated visualization is influenced by several medical and technical factors, including the patient's position during the procedures, such as TBE performed while standing versus endoscopy performed while lying down. Additionally, discrepancies may arise when captures are taken on different days, as the esophagus may (slightly) shift or change position over time. However, this is deemed an acceptable approximation from a medical viewpoint, and such movements of the esophagus are deemed negligible. A critical factor from an implementation view-point is the pixel-to-centimeter ratio derived from HRM and TBE data. It is important to note that endoscopic images lack a predefined pixel-to-centimeter ratio. Therefore, the TBE ratio is employed instead. To validate this method, we created reconstructions both with and without endoscopy data. Next, we compared the calculated metrics, which result in similar outcomes, confirming reasonable accuracy. The pixel-to-centimeter ratio determined using the, in the TBE annotated HRM sensor positions, is combined with manufacturer-provided information about the sensor distances in cm, and applied for subsequent metric calculations. Further evaluating reconstruction quality, additional tests were conducted, comparing our reconstruction algorithm against CT-based 3D-reconstructions.[16] At the time of writing, this research module is in its final stages. Both methods appear to yield comparable results (e.g. for the calculated indexes) that are considered medically acceptable, while 100% overlap of both reconstructions is not possible. For instance, CT scans are captured while the patient is lying down in contrast to TBE. We aim for medical applicability of the approximated reconstruction, and these measures are intended to assess and (continuously) support desired, and realistic outcomes through the interplay of medical and technical insights on the novel data type. Overall, we are convinced, that *EsophagusVisualization* designs well approximated 3D-reconstructions that are medically trustworthy.

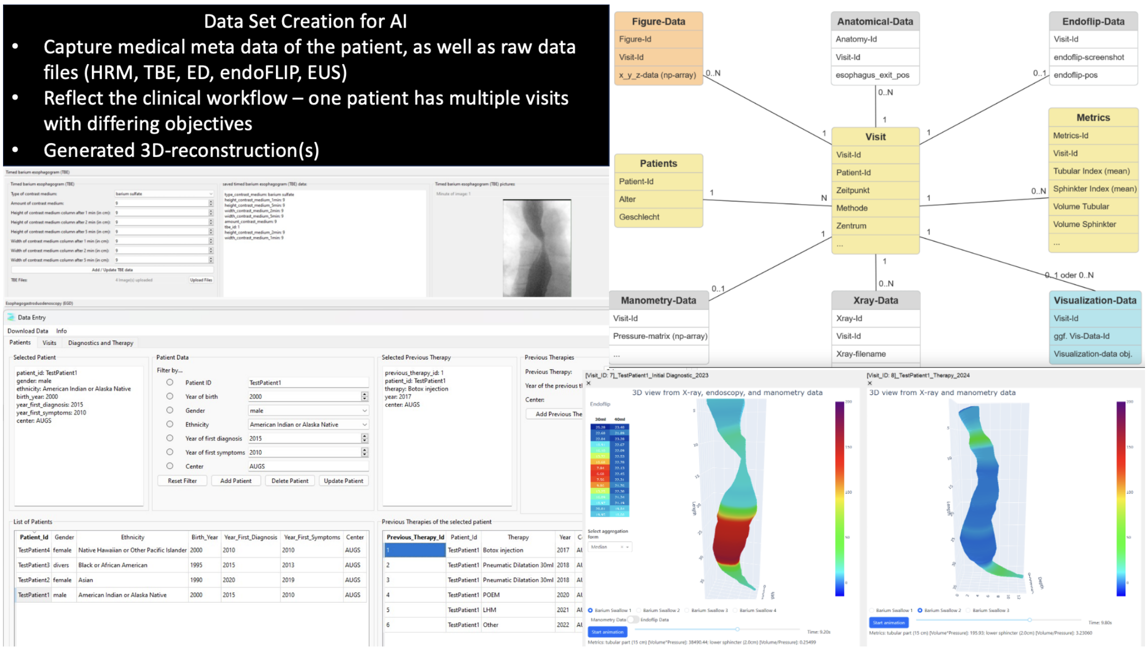
In summary, assessing the 3D-reconstruction, and the described workflow centered around the human-in-the-loop are crucial prerequisites to building a reliable data set based on this novel, multi-modal data type. This enables AI to be applied for different scenarios, as pointed out by our long-term project vision. Overall, this combination of (tested) automation and human oversight is essential for achieving reliable results in clinical applications that ultimately enhance patient well-being.

5.2.2.3. Data Set Creation & Application of AI

This section explores our long-term vision in detail, with respect to data set creation, and ideas/tasks towards intelligent augmentation centered around *EsophagusVisualization*. AI enhancements comprise implemented technical improvements of the inner-software workflow, envisioned support of the medical workflow through AI-generated 3D-reconstructions, as well as CDSS regarding treatment choices. Challenges emerge around compliant data set creation, as well as reliable AI output evaluation for more complex scenarios, including ethical questions, and the required quality to enhance patient well-being.

Multi-functional Data Set Centered around the previously introduced, multi-modal 3D-reconstruction of the esophagus, and aiming to support medical re-

[16] CT scans are ethically not permissible for Achalasia patients, but we had a case where, for diagnosis, all Achalasia data samples of a patient were additionally accessible.



Data Set Creation for AI

- Capture medical meta data of the patient, as well as raw data files (HRM, TBE, ED, endoFLIP, EUS)
- Reflect the clinical workflow – one patient has multiple visits with differing objectives
- Generated 3D-reconstruction(s)

Amount of contrast medium: 9
Height of contrast medium column after 1 min (in cm): 9
Height of contrast medium column after 2 min (in cm): 9
Height of contrast medium column after 5 min (in cm): 9
Width of contrast medium column after 1 min (in cm): 9
Width of contrast medium column after 2 min (in cm): 9
Width of contrast medium column after 5 min (in cm): 9

height_contrast_medium_1min: 9
height_contrast_medium_5min: 9
width_contrast_medium_2min: 9
width_contrast_medium_5min: 9
amount_contrast_medium: 9
tbe_id: 1
height_contrast_medium_2min: 9
width_contrast_medium_1min: 9

TBE Files: 4 image(s) uploaded

Selected Patient

patient_id: TestPatient1
gender: male
ethnicity: American Indian or Alaska Native
birth_year: 2000
year_first_diagnosis: 2015
year_first_symptoms: 2010
center: AUGS

Patient Data — Filter by...

- ○ Patient ID: TestPatient1
- ○ Year of birth: 2000
- ○ Gender: male
- ○ Ethnicity: American Indian or Alaska Native
- ○ Year of first diagnosis: 2015
- ○ Year of first symptoms: 2010
- ○ Center: AUGS

Selected Previous Therapy

previous_therapy_id: 1
patient_id: TestPatient1
therapy: Botox injection
year: 2017
center: AUGS

List of Patients

Patient_Id	Gender	Ethnicity	Birth_Year	Year_First_Diagnosis	Year_First_Symptoms	Center
TestPatient4	female	Native Hawaiian or Other Pacific Islander	2000	2010	2010	AUGS
TestPatient3	divers	Black or African American	1995	2015	2013	AUGS
TestPatient2	female	Asian	1990	2020	2019	AUGS
TestPatient1	male	American Indian or Alaska Native	2000	2015	2010	AUGS

Previous Therapies of the selected patient

Previous_Therapy_Id	Patient_Id	Therapy	Year	C
1	TestPatient1	Botox injection	2017	Al
2	TestPatient1	Pneumatic Dilatation 30ml	2018	Al
3	TestPatient1	Pneumatic Dilatation 30ml	2018	Al
4	TestPatient1	POEM	2020	Al
5	TestPatient1	LHM	2021	Al
6	TestPatient1	Other	2022	Al

Figure 5.16.: The current version of the *EsophagusVisualization* software enables the creation of a comprehensive data set. (source: based on *EsophagusVisualization* software)

search on Achalasia, as well as enhancement of diagnosis and treatment of the rare disease, multiple requirements for data set creation emerge. From a medical viewpoint, a lot of knowledge is currently unknown, therefore, a comprehensive collection of relevant factors, including patient data needs to be captured to enable a fruitful, data-driven analysis. In addition, the clinical workflow is characterized by multiple visits per patient that produce data for diagnosis, or monitoring of the esophagus' evolution before, and after therapy. Figure 5.16 summarizes the three main topics of the data set, comprising patient meta-data, medical information, capturing of and alignment with the clinical workflow, as well as the novel data type, including related information such as the calculated indexes. Further, since long-term, the data set is intended to be accessible by a number of different Achalasia centers worldwide, and patients may consult multiple centers during diagnosis, information on data sources is additionally included. Overall, the data set is envisioned to enable a comprehensive data-driven analysis centered around the application of AI. For this vision to become successful, first, a sufficient number of data samples, including 3D-reconstructions, needs to be acquired.

Data Base Implementation The current state of the prototypical *EsophagusVisualization* software (v3) allows for the creation of a comprehensive, multi-dimensional data set that is envisioned to long-term enable the, in the following introduced AI scenarios. The prototype integrates a PostgreSQL database[17],

[17] https://www.postgresql.org/

including a pgAdmin management interface[18], which is setup as two distinct services using Docker[19]. PostgreSQL was chosen as the database management system based on its open-source nature, ability to handle large and complex[20] datasets, queries [21], as well as scalability[22], which ensures adaptability for future expansion.[23] Due to the complex nature of the clinical workflow surrounding Achalasia, a SQL database was chosen over a NoSQL setup, even though it is characterized by unstructured and multi-modal data, which NoSQL is designed for. Overall, within our project, a structured schema is needed, that ensures relational integrity and data consistency, as well as enables complex data queries across e.g. multiple visits and patients, which are important requirements. [KW23] [MA21a] Overall, this containerized setup enables a lightweight and efficient way to run services without requiring traditional server infrastructure. As a result, for now, the prototypical database environment is simulated locally on the same hardware that runs *EsophagusVisualization*.

Challenges Drawbacks of this approach are centered around multi-user accessibility to allow a combination of physicians execute the data set creation simultaneously from distributed hardware. Currently, a seamless merging of generated data within a comprehensive data set for further analysis is not possible due to primary key conflicts when merging decentrally generated data bases. However, since these decentral databases share the same schematic structure, it is possible to address primary key conflicts as part of future work, for instance through the implementation of universally unique identifiers (UUID)[24]. Ideally, one day, the data set is hosted centrally by a competent organization that is equipped to handle technical challenges, which arise regarding the multi-modality and the increasing volume of data, as well as ensuring long-term support for the database and infrastructure. These issues demand robust solutions to handle diverse data formats and sources effectively, while also addressing scalability and sustainability to accommodate future growth and evolving requirements. Reliable infrastructure and database management are critical to ensure seamless integration, efficient processing, and long-term usability of the system. In addition, related challenges for data set creation towards translationality emerge, centered around

[18] `https://www.pgadmin.org/`

[19] `https://www.docker.com/`

[20] `https://www.postgresql.org/docs/current/largeobjects.html`

[21] `https://www.postgresql.org/docs/current/planner-optimizer.html`

[22] `https://www.postgresql.org/docs/current/limits.html`

[23] The Research Electronic Data Capture (REDCap) database management system is commonly used in medical translational research [MA15, 47]. Therefore, it was tested for applicability within our use case. Its flexible data upload options via a user-friendly web interface, enable integration with other systems like our *EsophagusVisualization* software. Testing confirmed successful data transfer from our system to REDCap. However, its rigid Entity-Attribute-Value data structure and lack of relational model support [MA15, 52] hinder efficient data retrieval and nested structures. In addition, since it is primarily designed for clinical surveys and trials [MA15, 48], it also struggles with unstructured data like images and videos, making it unsuitable for our project's needs.

[24] `https://www.postgresql.org/docs/current/datatype-uuid.html`

regulatory requirements and intricate ethical questions regarding e.g. patient privacy rights and data governance. In the best case, the data base is globally accessible to access and deposit a growing number of Achalasia patients, since it is a rare disease, and data scarcity is an issue. Ultimately, these considerations form the foundation for the responsible use of AI.

Segmentation The primary goal of this AI application scenario is to accelerate the creation of 3D-reconstruction data. Therefore, the segmentation model plays a key role in improving the previously introduced *EsophagusVisualization* workflow for creating 3D-reconstructions. It enables the automatic extraction of predefined shapes from clinical raw data, which is a relevant step for TBE, and endoscopic images. While numerous models and architectures exist for image segmentation [ER24], our challenge lies in the limited availability of labeled raw data, which significantly impacts the performance and reliability of these models. Within the software's workflow, the medical user already annotates the TBE and endoscopic images for shape extraction. Therefore, the current prototype (v3) integrates (automatic) multi-label raw data annotation export for segmentation data set creation, which is locally stored. The esophagus' shape, as approved by the physician, is automatically exported. Additionally, within the TBE image, the user can select to annotate the skeleton, and/or contrast agent for further experimentation. Currently, v3 includes a TBE-segmentation model based on nnU-Net[25], and the model selection process is outlined in more detail in chapter 8, illustrating MQG4A. This AI application scenario is deemed low-risk, since the medical user automatically evaluates, and possibly, adjusts the segmented shape as part of the *EsophagusVisualization* human-in-the-loop workflow. Regarding practicality, the trained model requires GPU power for shape estimations. Therefore, the medical user needs a high-performance technical setup. Otherwise, shape segmentation can take several minutes. Finally, this approach to TBE segmentation is transferable to endoscopic images, as part of future work.

Outlook: Generative AI GenAI has the potential to significantly support the multi-modal clinical workflow by enabling the intelligent optimization of 3D-reconstruction data creation. Ultimately, we aim for real-time generation of 3D-reconstructions during endoscopic procedures, enhancing the efficiency and precision of medical interventions, as well as to relieve the medical system. Currently, there is extensive research and numerous publications focused on 3D modeling and AI in medicine [RM24] [SM23] [ZY24]. However, specific challenges remain, particularly regarding 3D-reconstructions of the esophagus, and generally, ensuring the accuracy of these models, which is crucial to enhance patient well-being. As a result, a robust validation process of the generated 3D-reconstructions is necessary, with accuracy being continuously evaluated throughout the clinical workflow. This could continue previously introduced considerations, and indexes. Generally, it is essential to determine which medical and technical information is required to ensure both the reliability of the 3D-reconstructions and

[25] `https://github.com/MIC-DKFZ/nnUNet`

their seamless integration into clinical practice. Therefore, a next step of the project comprises growing a GenAI model that, based on endoscopic images (or, possibly videos) is capable to generate 3D-reconstructions. A fitting model needs to be selected that is capable to learn the relationship between the esophagus' shape and pressure values in relation to information extracted during the endoscopy. The novel data type, which is based on multi-modal medical data and medical knowledge, functions as groundtruth for that model. Consequently, the data set needs to be created prior/in parallel to model development, with a growing number of 3D-reconstructions. This may include multiple reconstructions per patient, an information that is possibly relevant for the model. More profound considerations regarding model architecture and lifecycle setup are part of future work.

Outlook: Clinical Decision Support CDSS systems play a central role in technical support for the provision of well-structured and meaningful therapy recommendations tailored to individual patients. The primary objective is to enable data-driven insights on treatment choices through effective human-machine teaming, where clinical expertise and AI work together to improve decision-making through combining data-driven findings with practical human experience. While there is extensive research and numerous publications on CDSS, even in the context of the esophagus (no Achalasia-specific CDSS exists to date), significant challenges remain [MM24a] [RC22] [DL22b]. These include ensuring the system is aligned with specific medical use cases, while addressing concerns about black-box behavior, and building trust among clinicians (and patients). Therefore, a rigorous validation process is essential, focusing on evaluating therapy outcomes within the medical context of Achalasia, as previously outlined, to ensure reliability and effectiveness. To achieve these goals, it is critical to identify the medical and technical information necessary to ensure the success of CDSS implementations, fostering both usability and clinical accuracy surrounding Achalasia. A precision of these considerations is part of future work, while we aim to capture all relevant data early on within the previously introduced multi-modal data set.

After outlining the medical domain with respect to AI, and introducing two different (fictional) medical use cases, the next part evaluates MQG4AI interaction scenarios. First, in chapter 6, a reliable and generic explanation lifecycle stage is designed, including the assignment of a technical guideline for explanation evaluation to MQG4DK. This demonstrates the leaf-QG template and MQG4DK contribution of design decision-making, as well as information linking with contextual information blocks, highlighting the role of risks and best practices. Next, the following chapter 7 contributes a leaf-QG compilation to MQG4DK, aiming to extract a reliable metrics selection workflow based on the previously introduced fictional ECG use case. Finally, MQG4A is illustrated through different template-versions that reflect design decision-making for segmentation model selection in the context of the *EsophagusVisualization* software for Achalasia in chapter 8.

Part III.

APPLICATION AND EVALUATION

6 MQG4DK Design – Explanations during Development & Leaf-QG Information Processing

After introducing MQG4AI in the previous chapters, we present a generic explanation stage design, illustrated on GitHub[1], and based on [EM25a] in this chapter. We align XAI best practices within the AI development phase, which is critical for ensuring transparency, traceability, and risk-awareness in AI systems, as introduced in section 2.2.2. We aim towards quality by AI lifecycle design. While this chapter focuses on the explanation stage within model development, the design approach is in principle transferable and adaptable to other lifecycle phases. We illustrate MQG4AI interaction for MQG4DK, and demonstrate a possible workflow to support the design of the generic AI lifecycle, including the documentation of AI risks, as exemplified in section 6.1 for XAI. Quality is evaluated through alignment with the *IEEE Guide for an Architectural Framework for Explainable Artificial Intelligence* [Art24] [EM25a, 12] in section 6.2. Finally, the proposed MQG4DK design workflow includes an illustration of the leaf-QG layout, as introduced in section 4.3. We outline a use case-independent, technical guideline contribution to the design knowledge base (MQG4DK) for evaluating explanation quality focusing on Local Interpretable Model-agnostic explanations (LIME) and SHapley additive exPlanations (SHAP) [LGC24] in section 6.2.2. Consequently, in case MQG4A versions apply LIME and/or SHAP explanations, this design guideline is pulled from MQG4DK.

Building on this, the next chapter 7 elaborates on a more complex MQG4DK contribution in the form of a compilation of leaf-QGs, addressing aspects such as inter-QG relations. Additionally, this contribution relates to a fictional use case centered around multi-label ECG classification [EM24], as introduced in section 5.2.1, highlighting the inclusion of domain and AI system-specific information. Finally, chapter 8, describes a possible simulation of how MQG4A scenarios could unfold, based on a retrospective analysis of segmentation model selection for the tool *EsophagusVisualization* centered around the rare disease Achalasia, as intro-

[1] `https://github.com/miriamelia/MQG4AI/blob/main/README.md`

duced in 5.2.2, before concluding our contribution in chapter 9. In summary, the following chapters provide an overview on how to interact with MQG4AI to concretize the, in part II introduced blueprint, highlighting RAI and MQG4AI's building blocks centered around generic and customizable *Quality Gates* (QG).

6.1. Evaluating Explanations – Risks & Best Practices

We intend to outline a generalizable MQG4AI design workflow that enables defining the generic AI lifecylce based on best practices that mitigate risks by design, in a decentralized manner (MQG4DK). This DSR mechanism within MQG4AI is further detailed in section 3.3.7. Concretely, we exemplify the process in a modularized manner for the generic *Explanation* lifecycle stage. As introduced in section 2.2.2, the rise of black-box models is linked to a rapidly increasing interest in model interpretability [HS20, 3], which results in the application of XAI methods along the AI lifecycle, including their translation to relevant stakeholders, as well as quality criteria for "good" explanations. This section introduces two key risks associated with model opacity, reflecting concerns related to explanation output and the user, as well as the question of suitability regarding XAI method application. Namely, in the following, we highlight *Incomprehensible Explanations to the User* and *Unfaithful or Unreliable Explanations*, as well as best practices for mitigation. Both identified AI risks are assigned to *Explainability* as subsection of *Transparency* within the MQG4AI *Risk Management* information block on GitHub.[2] Rounding up our contribution, the next section 6.2 provides a detailed view of the leaf-QG information processing template, as introduced in section 4.3. Concretely, we illustrate how to append a technical guideline in form of an existing publication [LGC24] to MQG4DK [EM25a]. Finally, section 6.3 evaluates the proposed approach.

6.1.1. Explanation Output & the User

Explanations only then fulfill their intended purpose, when they are interpreted as intended by their target audience, as detailed in section 2.2.2.2. This section focuses specifically on the role of the AI system user, who, unlike developers, is typically the recipient of the explanation without possessing extensive technical knowledge. Consequently, the explanation output must be tailored to the needs of this audience, requiring additional layers of information transformation to ensure clarity, relevance, and interpretability. Ensuring effective user compre-

[2] https://github.com/miriamelia/MQG4AI/tree/main/MQG4DesignKnowledge/3_RiskManagement/AI_Risks/4_Transparency/Explainability

hension is critical to achieving the intended use and trustworthiness of the AI system.

6.1.1.1. Risk: Incomprehensible Explanations to the User

While research on model interpretability has expanded rapidly across fields, little is known about how practitioners perceive and implement interpretability within their workflows. This gap in understanding may hinder interpretability research from addressing critical real-world needs or result in impractical solutions. [HS20, 1] The risk *Incomprehensible Explanations to the User* occurs when the explanation methods and their presentation are not aligned with the user's ability to comprehend, leading to confusion and misunderstanding. Explanations that are overly vague, too complex, or poorly structured can hinder the user's understanding of how the model reaches its conclusions and how to apply the results in a specific context. This, in turn, can erode trust in the system and result in various misuse scenarios depending on the concrete use case at hand. Therefore, considering the "human-AI interface" [AS23, 32] plays a crucial role when designing explanations along the AI lifecycle.

6.1.1.2. Towards Comprehensible Explanations

User comprehensibility comprises a necessary quality criteria of explanations, especially in healthcare [BS23, 19], and needs to be considered from the start of explanation method design within a particular project. In section 2.2.2.5, we introduce considerations surrounding explanation quality criteria and address how to organize relevant information for reliable explanation design in section 2.2.2.3. Overall, to ensure a trustworthy application, explanations must be communicated clearly to users, which needs to be evaluated. To begin with, we outline the following best practices for designing the explanation stage with a focus on the AI system user:

- Develop informative interfaces that accompany the raw explanation output at varying levels of complexity to enhance user interpretation.

- A well-structured user interaction flow is essential to facilitate understanding of the system's functionality, including its context-specific operation and limitations.

- A notification protocol should also be in place to inform users of relevant updates that may impact the AI system's behavior.

Ultimately, the concrete implementation of such measures must be tailored to the specific use case. For example, in an intelligent CDSS, a pop-up notification could alert physicians to potential inaccuracies when the system processes data from a female patient, if the model was primarily trained on male patient data.

6.1.2. Explanation Application & Suitability

In addition to adjusting explanation output to different stakeholders, it is equally important that developers understand "[...] the general behavior of explanation methods [...]" [HA23, 14] to avoid mistakes when applying explanations and choosing methods for the use case at hand [HA23, 14]. The following example aims to clarify complexities of choosing fitting explanations for LIME and SHAP explanations, which are introduced in section 2.2.2.3 and correspond to the blue explanation classification in Figure 6.2:

- *LIME* can identify which features play a critical role in the model's decision-making process. However, LIME's reliance on random perturbations leads to inconsistencies. Additionally, LIME assumes a linear relationship in local data, which may not always hold, increasing the risk of misinterpretation. The method also depends on strong assumptions of feature independence, making it difficult to define appropriate similarity measurements for assigning weights. [BS23, 5]

- *SHAP* is designed to evaluate the contribution of each feature, leveraging game theory for fair predictions across distributed feature values. However, this can sometimes lead to conflicting explanations when comparing different predictions. In combination with LIME, SHAP enhances understanding of both methods. A key advantage is that SHAP ensures consistency between global and local interpretations, unlike approaches that e.g. mix LIME with other global interpretability methods. However, SHAP is computationally expensive for large datasets, rendering the method too time-consuming for complex and large data. [BS23, 4]

In summary, XAI methods need to be evaluated carefully for applicability to individual use cases, including the identification of trade-offs. In addition to profound considerations surrounding explanation design, this may include complementing methods to quantitatively evaluate explanations through the calculation of metrics, as well as human-based evaluations for comprehensibility assessment.

6.1.2.1. Risk: Unfaithful or Unreliable Explanations

XAI pitfalls [dH22] exist, and several issues reported in the literature highlight both technical limitations and the previously addressed user misinterpretations [BC24]. Consequently explanation method selection is a non-trivial design decision and we introduce an exemplified selection process in section 2.2.2.4. The risk *Unfaithful or Unreliable Explanations* occurs when explanation methods fail to accurately represent a model's true reasoning, leading to misunderstandings of its decisions. Such explanations can mislead stakeholders, potentially resulting in harmful actions during the model's lifecycle and diminishing trust in the system during application.

6.1.2.2. Preferred Use of Interpretable Models (Ante-hoc)

The best way to mitigate the risk *Unfaithful or Unreliable Explanations* is to use ante-hoc explainability methods, where applicable for the specific use case. Ante-hoc models are inherently interpretable and designed to provide clear insights into their decision-making processes without relying on post-hoc explanation techniques. These models can range from simpler ones, like decision trees, to more complex ones with interpretability enhancements, such as DNNs with prototype layers, or other symbolic methods [BB24] [SZ21] that include relevant domain knowledge with network design. A neural network with a prototype layer classifies inputs by comparing them to representative prototypes, ensuring predictions are grounded in human-understandable references [LO17]. By incorporating interpretability directly into the model architecture, ante-hoc models reduce the risk of unreliable explanations, as the explanations are naturally aligned with the model's decision-making process. As a result, evaluation primarily focuses on model performance. However, this may not always be a feasible or reasonable approach, and post-hoc methods tend to be applied more commonly [VL20, 16].

6.1.2.3. Validate the Explanations

Emphasizing post-hoc explanations, it is crucial to validate explanations in terms of their desired mathematical properties and their relation to the user's understanding to effectively address the risk *Unfaithful or Unreliable Explanations*. This ensures that explanations are not arbitrary or misleading, but rather meaningful, consistent, and trustworthy. Additionally, the desired properties of an explanation depend on the specific technique used. In section 2.2.2.5, we introduce desirable quality criteria for explanations in more detail, further highlighting the design decision's complexity.

In summary, effective explanations should adhere to key principles to ensure clarity and usefulness, while addressing technical intricacies along the AI lifecycle and its evolutions. In the next section, we attempt to translate these best practices into generalizable design decisions along the AI lifecycle.

6.2. Towards a Generic & Reliable Explanation Stage

After identifying relevant criteria for integrating explanations into the AI lifecycle, we design the generic *Explanation* stage, with a focus on organizing reliable implementation concepts that align relevant design decisions with individual use case-specific methods. Namely, we derive three QGs, as demonstrated in section 4.2.3.4. Next, we present the leaf-QG template, as introduced in section 4.3. Concretely, we outline a concrete example contribution to MQG4DK on how to evaluate the quality of LIME and SHAP applications.

6.2.1. Proposed Explanation Lifecycle-Stage QG Structure

To bridge the gap between use case specificity and broad applicability, we propose a combination of collection- and leaf-QGs that are designed to be adaptable across various use cases for MQG4A scenarios. Simultaneously, the proposed design offers a structure for assigning use case-specific guidelines at lower hierarchical levels to MQG4DK. The resulting *Explanation* lifecycle stage comprises three primary stages: *Configuration*, *Evaluation*, and *User Interaction*, as depicted in Figure 6.1.[3]

6.2.1.1. *Collection-QG* User Interpretation

Collection-QG User-Interaction ensures that explanations are presented effectively and communicated clearly to relevant stakeholders, shedding light on the com-

[3] The structure is similar to the proposed sub-level for the *Development* lifecycle stage (Configuration, Evaluation, Optimization, Explanation), as outlined in section 4.2.3. However, while explanation method configuration is defined as a leaf-QG, model configuration involves more intricate sub-steps, thus making it a collection-QG in our proposition, focusing on MQG4DK. In the context of MQG4A, this differentiation depends on the thoroughness of design decisions and the extent to which established, possibly codified, methods are considered. Additionally, explanation *Optimization* is conducted through various MQG4A-template versions, addressing different input information interdependencies, such as data pre-processing, and the underlying model that shape explanation performance. There are no specific explanation optimization methods, in contrast to model *Optimization*, which involves individual sub-steps, such as post-processing methods. These template versions are simulated in chapter 8.

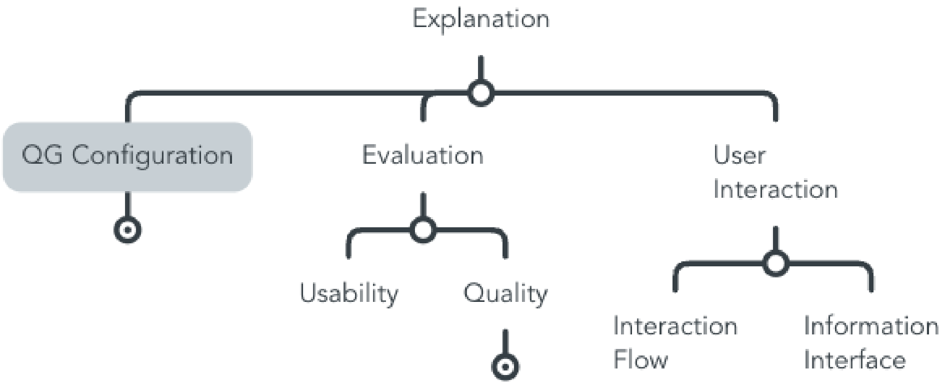


Figure 6.1.: Three identified design decisions that comprise the generic *Explanation* stage. In this proposition, focusing on MQG4DK, only QG Configuration comprises a leaf-QG, the rest are collection-QGs.

plexities of the decision-making process behind specific AI output. As depicted in Figure 6.1, based on the previously introduced best practices centered around comprehensible explanations, we derive two sub-QGs: a well-defined explanation *QG Interaction Flow* once deployed in the intended real world setting, as well as an accompanying *QG Information Interface* to communicate complementary information that enhances the recipient's interpretation of the explanation's message. Overall, this stage emphasizes the presentation of the generated explanations, leveraging user studies and *Usability* testing to gather actionable insights into user interactions, aligned with explanation *Usability* evaluation, as introduced later, with respect to explanation assessment. Originally, *Usability* is derived from *user friendly* [ARD09, 2] and, in alignment with technological advancements, the concept has evolved and became more complex to implement. A comprehensive analysis and taxonomy of *Usability* in the context of software development is provided in [ARD09], for example. Possibly, the introduced detailed organization of related sub-concepts [ARD09, 5] can be consulted as a guideline for design and evaluation of *collection-QG User-Interaction*, complementing the, in the previous section introduced, structure to organize explanations. Focusing on the human recipient, the overarching goal is to develop well-designed GUIs, so that explanations enhance human-machine interaction towards ensuring the intended use. Concrete methods for implementation of these *Explanation* design decisions are beyond the scope of our current analysis.

6.2.1.2. *Leaf-QG* Method Configuration

Leaf-QG Method Configuration refers to a summary of relevant metadata and explanation characteristics that apply to any explanation method. These characteristics define the setting for each XAI method, and multiple methods can be incorpo-

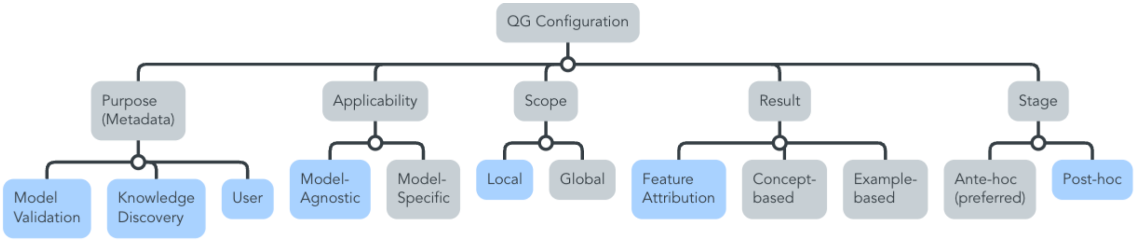


Figure 6.2.: The proposed setup for explanation method leaf-QG Configuration. For example, if the use case requires feature importance (result) and local (scope) explanations, methods like SHAP or LIME can be selected. Additionally, both SHAP and LIME explanations can be used for purposes like model validation, model discovery, or user interpretation (purpose), are model-agnostic (applicability), and are post-hoc (stage). These options are highlighted in blue and comprise the suitable setting for the concrete design decision contribution to evaluate LIME and SHAP explanations in the next section 6.2.2.

rated. As illustrated in Figure 6.2, the definition of XAI methods includes factors such as purpose, applicability, scope, result, and stage, with various options specified for each. Ideally, these options cover all potential explanation methods, forming a comprehensive framework to guide the selection of the most appropriate approach:

- *Purpose* refers to the reason why explanations are generated, which is crucial for risk mitigation. Developers can annotate whether explanations are needed to validate the model, assess data preprocessing techniques, inform stakeholders about model decisions, or uncover new insights learned by the model. This information aids in selecting the appropriate evaluation strategy for the generated explanations.

- *Applicability* indicates whether the explanation method is *model-agnostic*, therefore, applicable to any ML model, or *model-specific*, meaning it is designed for a particular type of model. This decision influences both computational resource requirements and model flexibility. Model-specific methods may demand more computational power during implementation but could be more efficient once deployed, while model-agnostic methods are more flexible but potentially more resource-intensive.

- *Scope* defines whether an explanation is *global* and thus covering the entire model or *local*, focusing on individual predictions. The scope should be chosen in alignment with the level of detail required for the specific audience or task.

- *Result* refers to the format in which the explanation is presented, such as text, visualizations, or statistical summaries. This format should be tailored to meet the needs of the target audience. For example, in healthcare, a local explanation might take the form of a visual heatmap that highlights the areas of an image influencing a diagnosis, offering actionable insights for medical professionals.

- *Stage* describes the timing and process by which explanations are generated in relation to model development. It specifies whether the explanations are derived from an already trained model (post-hoc) or if modifications to the model are necessary (ante-hoc). This choice impacts processes like evaluation and user interpretation. Generally, post-hoc explanations require more rigorous validation to confirm their reliability and relevance, in contrast to ante-hoc methods that are inherently (made) explainable.

These characteristics should be defined based on the use case-specific requirements, such as the model's explainability needs and existing best practices. Contextual factors, such as computational resources, model complexity, and the interpretability needs of stakeholders, equally influence XAI method configuration.

6.2.1.3. *Collection-QG* Method Evaluation

Collection-QG Method Evaluation highlights the importance of evaluating explanations in terms of both, *Usability* and *Quality*, which comprise overarching criteria to evaluate explanations, as shown in Figure 6.3. Generally, the evaluation process presents challenges due to domain-specific limitations, the wide array of properties that can be assessed, and difficulties integrating case-specific metrics. The lack of universal methods for automating this evaluation poses further complications [AMJ18], and "[...] the quantitative assessment of neural network explanations remains non-trivial" [HA23, 5]. Despite these obstacles, *collection-QG Method Evaluation* lays the groundwork for a thorough and meaningful evaluation process, ensuring explanations are reliable, aligned with the use case's goals, and that identified risks are addressed appropriately:

- *Usability*, or subjective evaluation, focuses on how understandable and interpretable the explanation is for its intended audience. It includes factors like user satisfaction, ease of comprehension, and the level of trust and transparency fostered by the explanation. These factors are usually assessed through qualitative methods such as surveys, case studies, and focus groups. Usability can also be evaluated in terms of conciseness (parsimony), comprehensiveness (coverage), and clarity (overlap) [LH16]. Additional criteria may follow existing taxonomies for explanations [HS20, 7] and usability in general [ARD09]. Finally, usability evaluation is closely linked with *collection-QG User Interpretation* and evaluates how well implemented methods perform.

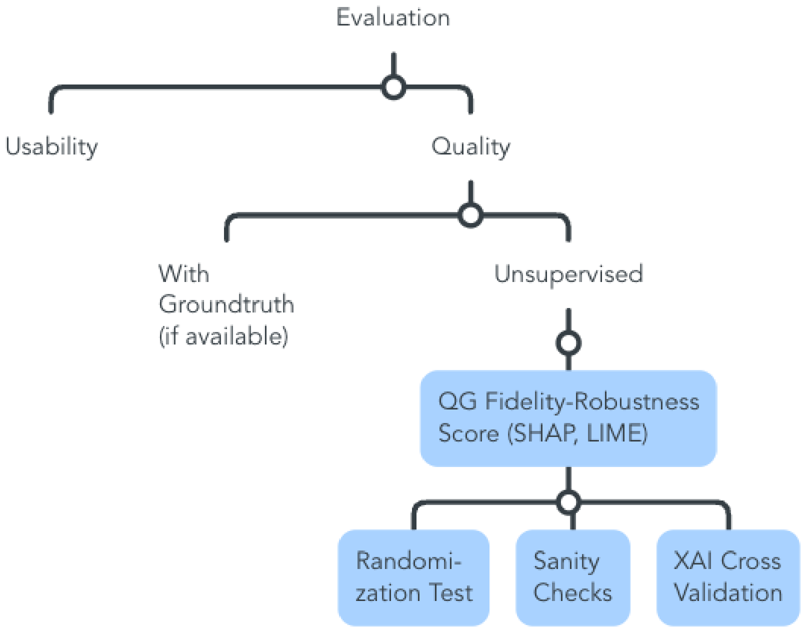


Figure 6.3.: The proposed generic structure to organize explanation evaluation information, including a proposition for unsupervised quality assessment, as introduced in the next section 6.2.2.

- *Quality*, or objective evaluation, focuses on the explanation's accuracy and reliability. As shown in Figure 6.3, quality can be assessed using the explanation's ground truth or in an unsupervised manner. If the ground truth is not accessible, mathematical properties can be identified and assessed using computational techniques, to provide a quantitative evaluation of explanation reliability. These properties serve as foundational elements for a robust evaluation framework. Our exemplified leaf-QG contribution is assigned to this *collection-QG Unsupervised Explanation Quality Evaluation*. As introduced in the next section, two key interrelated properties are identified (robustness and fidelity), including computational methods to evaluate SHAP and LIME explanations across these dimensions. *Robustness* measures how consistent an explanation remains when small variations are introduced to the input data, while *fidelity* quantifies how accurately the explanation reflects the model's true decision-making process. [LGC24] Another approach is presented in [HA23], where the authors focus on two key aspects: resilience to noise and reactivity to randomness, equally enabling quality assessment without relying on ground truth labels.

In summary, the *Explanation* stage in MQG4AI integrates explainability-related design decisions into the AI lifecycle via collection- and leaf-QGs, guided by best practices towards risk mitigation. To operationalize this stage, further leaf-QGs can detail specific methods, as illustrated in the next section with an evaluation of LIME and SHAP explanations as a contribution to MQG4DK.

6.2.2. *Leaf-QG* Creation – Exemplified for Quality Assessment of LIME & SHAP Explanations

After outlining the generic *Explanation* lifecycle stage, which consists of the three generic sub-design decisions *Configuration*, *Evaluation*, and *User Interaction*, we append a technical guideline for assessing the quality of LIME and SHAP explanations to MQG4DK in this section. As depicted in Figure 6.3, it is classified as an *unsupervised* quality evaluation method. Concretely, we introduce *leaf-QG_FidelityRobustnessScore_(SHAPLIME)*[4], which builds on insights from [LGC24] and details how to apply unsupervised methods to evaluate explanation quality through a combined *fidelity* and *robustness* evaluation score. This leaf-QG is pulled from MQG4DK, if LIME and/or SHAP explanations are relevant to the individual project, which is configured to define MQG4A-v0. The envisioned process to collect and apply RAI knowledge within MQG4AI is introduced in sections 3.3.7 and 4.2.1, as well as exemplified in chapter 8.

- **Robustness** assesses the stability of an explanation by evaluating its consistency when minor changes are made to the input data. Robustness is crucial for both trustworthy models and interpretable explanations. Stable interpretation methods should remain consistent despite minor input changes, enhancing persuasiveness and trust. Low stability undermines reliability, making explanations seem accidental. Therefore, explanation robustness, or stability, should be quantified to ensure interpretability. [Art24, 17]

- **Fidelity** comprises a critical aspect of post-hoc interpretability, as it determines how well explanations align with the underlying model's decision-making process. High-fidelity explanations are essential for ensuring that interpretations are meaningful and trustworthy. Fidelity can be categorized into local and global fidelity. Local fidelity ensures that explanations accurately represent the model's predictions for a specific data point or a small group of instances, while global fidelity requires that explanations remain consistent across all instances. Importantly, strong local fidelity does not necessarily imply strong global fidelity. Ensuring high fidelity helps stakeholders assess the reliability and usefulness of explanations. [Art24, 17]

The following sections exemplify how leaf-QGs process information to reflect concrete lifecycle design decisions within MQG4DK, leveraging existing RAI knowledge in the form of a published technical guideline. Leaf-QGs are introduced in section 4.3 in more detail, and the following structure reflects the demonstrated information layers, as depicted in Figure 4.21.

[4] `https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/2_Lifecycle/2_Development/4_Model_Explanation/Method_Evaluation/Quality/QG_FidelityRobustnessScore_(SHAPLIME).md`

6.2.2.1. Naming & Tags

The collection-QG *Explanation Quality Evaluation* specifically addresses the risk of *Unfaithful Explanations*, as previously introduced in section 6.1.2.1, contributing to overall *Transparency*. This is where the *Fidelity-Robustness Score* [LGC24] is located along the AI lifecycle. Following the previously introduced **QG Naming Structure** (design decision's name; view on its applicability), *QG_FidelityRobustnessScore_(SHAPLIME)* introduces a Fidelity-Robustness Score to assess the quality of explanations that are generated by SHAP and LIME. Additional contributions to the collection-QG *Explanation Quality Evaluation* may incorporate other assessment methods [HA23] or extend the Fidelity-Robustness Score with further metrics, tailored to different XAI techniques, for instance. Related **QG Tags** within MQG4DK [WH22] aim to enable a use case-specific MQG4A-v0, and this particular realization of a generic design decision is included if applicable to the individual project's configuration. They are illustrated as follows:

- *Name*: QG_FidelityRobustnessScore_(SHAPLIME)
- *Intent*: Explanation method applicability evaluation
- *Problem*: Unfaithful explanations
- *Solution*: Explanation evaluation metrics
- *Applicability*: SHAP/LIME-generated explanations only
- *Consequences*: Fidelity-Robustness Score (quantification of required explanation qualities)
- *Usage Example*: None (a purely technical guideline)

The next leaf-QG information layer, the *QG Interdependency Graph*, represents connections with relevant *Input and Output Information*, derived from related lifecycle stages and supplementary information blocks. It highlights key interdependencies to ensure a structured integration of information along the AI lifecycle. Generally, we propose to design leaf-QGs based on the *Socratic Method* [IA12]. This approach incorporates a guided question-answer process that results in guiding questions (GQ) that support bridging the gap between generalizability and use case-specificity of RAI lifecycle design blueprint.

6.2.2.2. Input Information

Input Information addresses the question: *What information is necessary to execute the method and generate the content?* QG_FidelityRobustnessScore_(SHAPLIME) relies on SHAP and/or LIME explanations, which must be derived from the predictions of a previously trained model. Naturally, the model must support the generation of such explanations, which may impose constraints on model selection. Additionally, factors such as the chosen model, dataset, preprocessing steps, and overall model performance influence the quality of the resulting explanations. Therefore, relevant lifecycle implementation QGs include:

- *QG_Utilization_(Data)* provides insights into data quality through statistical analysis and preprocessing procedures.

- *QG_Configuration_(Development)* summarizes details on the model whose output is being explained.

- *QG_Evaluation_(Development)* encompasses performance metrics used to monitor the model, contribute information to the quality of explanations.

- *QG_MethodConfiguration_(Explanation)* specifies the applied explainability technique to be evaluated, as depicted in Figure 6.2.

Additionally, contextual *AI System Information*, such as its intended use, may be linked. However, no AI system-specific information was identified in this case.

6.2.2.3. Output Information

Output Information addresses the question: *Which stages are impacted and what additional information might be required?* It centers on gathering relevant information for post-market monitoring and further design decision-making along the AI lifecycle. The proposed QG assesses explanation quality by combining fidelity and robustness as key evaluation metrics. As noted by [LGC24], explanations that lack fidelity can mislead users, creating a false sense of understanding. To address this concern, the application-oriented contribution presents a unified quality score that evaluates both fidelity and robustness at once, ensuring a more thorough and balanced evaluation.

With respect to the **Post-market Monitoring Layer**, output information answers the question: *Are LIME/SHAP explanations appropriate for explaining the model?* This is evaluated on a scale from 0 to 1, with recalculation being triggered whenever the model is retrained on new data, which should be incorporated into the monitoring strategy, including the score's interpretation and possible guidelines to action. A score of 0 indicates that LIME/SHAP explanations are unsuitable for the task and should not be trusted, while a score of 1 means the explanations are perfectly suited for the task. Scores between 0 and 1 indicate varying degrees of appropriateness. As a general guideline, scores above 0.8 are considered acceptable, and scores above 0.9 are excellent. If the score is lower, it is recommended to regenerate the explanations for the ML task, potentially adjusting the model, data, or feature importance method used. The impacted generic lifecycle collection-QGs include:[5]

- *QG_MethodConfiguration_(Explanation)* may require adjustment if the generated explanations do not meet the necessary standards, as determined by e.g. the calculated evaluation metrics.

- *QG_Deployment* necessitates functioning infrastructures to validate the quality and accuracy of explanations, linked to post-market monitoring data to continuously evaluate real world performance.

- *QG_Maintenance* emphasizes on-going monitoring and evaluation, ensuring that improvements are implemented based on feedback during operation.

Building on identified interdependencies, *QG Creation* consists of three key information dimensions for design decision-making, including an *Evaluation* layer to assess the selected method: *Content*, which provides a detailed explanation of the design decision; *Method*, which outlines the decision-making process behind the design choice at hand; and *Representation*, focusing on information transformation to stakeholder roles that comprise the respective target audience. This structured approach aims to enhance quality by design, as it encourages a re-evaluation of each individual design decision.

6.2.2.4. Dimension 1: Content Definition

To assess fidelity and robustness, a series of sanity checks and measures to implement stability against distribution changes are used to calculate the score. It is important to specify any assumptions made by the method. This contribution is based on the assumption that, if explanations are robust to different test sets, they are also robust to distribution changes within the test set (as the test

[5] *Deployment* and *Maintenance* are not the focus of our contribution, so they are only referenced, and our exemplified information collection may be incomplete or not sufficiently detailed for seamless application. As RAI design knowledge advances, more specific process steps can be identified. This approach to MQG4DK equally reflects MQG4A template conceptualization of evolving lifecycle design, where further details on real world model integration may lead to gradual refinements of interdependencies with model development, as the project advances.

set can only change if its distribution changes). Furthermore, if explanations pass fidelity checks, the mean of the generated explanations is considered representative of all explanations. The fidelity evaluation produces a binary score (0 or 1), while the robustness evaluation yields a value between 0 and 1. These two scores are then combined through multiplication, which is documented within the proposed leaf-QG.

6.2.2.5. Dimension 2: Content Definition Method

Complementing a concise summary of the individual design decision, additional details on the concrete implementation of the proposed leaf-QG are provided. Within QG_FidelityRobustnessScore_(SHAPLIME), relevant considerations are summarized as follows:

Fidelity *Sanity checks* are carried out using a series of randomization tests, including model randomization and data randomization checks, to assess whether the explanations genuinely reflect the model and data [AJ18]. The *data randomization test* compares explanations generated from a model trained on original data (base scenario) to those generated from a model trained on data with randomized class labels. The *model parameter randomization test* compares explanations generated under the base scenario to those generated from a randomly initialized model. For both tests, feature importance explanations are considered similar if their feature importance rank is comparable. However, since less important features contribute less to the model's overall behavior, ensuring similarity in the rank of the most important features is critical. In line with [LGC24], the proposed leaf-QG uses the *Normalized Discounted Cumulative Gain (NDCG)* metric to quantify the similarity between explanations, as it prioritizes the correct ranking of higher-importance features.

Robustness Additionally, robustness is assessed by calculating the similarity of explanations generated from data with slightly different distributions. Explanations are generated for models that are trained and tested on different data splits that introduce small variations in the data distribution. Again, the *NDCG metric* is used to quantify the similarity between these explanations.

Final Score Both results are then combined through multiplication to obtain a final score. The resulting score automatically is 0 if fidelity is 0. If a certain degree of model fidelity exists (fidelity equals 1), and the score will reflect the explanation's robustness to changes in the data distribution.

Considering MQG4A for instance, this method results in multiple lifecycle template versions for explanation assessment execution. The calculated scores are based on different data distributions, including original and randomized labels for fidelity, as well as multiple versions with slightly different data distributions for robustness calculation. This application is part of future work.

6.2.2.6. Dimension 3: Content Representation

The final *QG Creation* dimension *Representation* focuses on ensuring that relevant information is provided to the right stakeholders at the appropriate time, and in a reasonable manner, aiming to facilitate effective communication and compliance. Refer to sections 3.3.2.1 and 3.3.3.5 for more information on stakeholders. Among others, identified global roles include:

- AI Experts and Data Scientists (active): These stakeholders are responsible for re-evaluating data preprocessing, model development, and/or explanation generation.
- Domain Experts (consulting): They are involved in validating explanations subjectively to ensure relevance and accuracy in the domain context.
- Regulators (passive): Their role is to verify that the AI system complies with current regulations.
- AI Users (passive): These individuals interpret model decisions and aim to understand the underlying decision-making process.

6.2.2.7. Evaluation Layer

The *Evaluation* layer highlights open questions related to a particular design decision, which may serve as contribution for evaluating residual risks. The following list evaluates the proposed fidelity-robustness score:

- *Why were the score metrics chosen?* Fidelity and robustness metrics were selected based on their frequent reference in the literature and their applicability across various types of explanations. [LGC24]
- *How well tested is the design decision-making method?* The score is evaluated on Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and Extreme Gradient Boosting (XGBoost) models across two different datasets. [LGC24]
- *What is the core assumption?* The key premise for application is that the model-agnostic SHAP and LIME methods generate feature importance ex-

planations. They highlight the significance of each input feature in model predictions. As a result, the calculated score focuses on the similarity of feature importance ranks, as less important features are assumed to be less informative about the model. The similarity of explanations (measured by the NDCG) is defined in terms of both the similarity of feature ranks (order of features by importance) and the actual feature importance values. [LGC24]

- *How significant is the impact of the domain?* The focus on feature importance ensures that the score remains independent of specific domains or use cases, enabling broad applicability. [LGC24]

Concluding the detailed illustration of the leaf-QG information processing template, *Additional Information* is illustrated for *Risk Management*, emphasizing information linking across the AI lifecycle.

6.2.2.8. Risk Management Layer

QG_FidelityRobustnessScore_(SHAPLIME) offers a method to evaluate the robustness and fidelity of model explanations, providing a quantifiable score to address the risk *Unfaithful Explanations*. As previously introduced, it is linked as a possible risk control to address risks associated with the TAI criteria *Transparency - Explainability* within the MQG4AI lifecycle blueprint, corresponding to the *Assessment List for Trustworthy AI* (ALTAI) [Hig20]. The practical integration of this information linkage into MQG4A template versions remains a focus for future work, while we simulate an example scenario in chapter 8.

6.3. Evaluation

The, in the previous sections introduced workflow, illustrated for the *explanation* phase during model development combines risks and best practices with the AI lifecycle to achieve quality by design. Overall, the outlined procedure is envisioned to be transferrable to other stages and design decisions. In addition to generic explanation design, we illustrate RAI knowledge organization and sharing for *QG_FidelityRobustnessScore_(SHAPLIME)*, as a contribution to MQG4DK in this chapter. As a result of the proposed design structure, explanation-related MQG4DK scenarios focus on collecting existing explanation configurations, evaluation approaches, as well as methods to enhance user interaction.

Generic & Customizable Explanation Stage The resulting *Explanation* stage, as depicted in Figures 6.1 6.2 6.3, is alignable with the *IEEE Guide for an Architectural*

Framework for Explainable Artificial Intelligence (XAI) [Art24], which is comprehensively detailed in the corresponding contribution [EM25a, 12]. In essence, the framework offers a structured approach to defining and evaluating explainability in AI systems, emphasizing the adoption of diverse XAI methodologies tailored to intended use cases. It integrates content type (what to explain), communication (how to explain), and stakeholders (who receives the explanation) with lifecycle considerations, spanning pre-modeling, modeling, and post-modeling stages. Furthermore, explanations are assessed through qualitative and quantitative usability measures, as well as subjective and objective metrics that mathematically evaluate explanation quality. We establish a correspondence between the proposed MQG4AI information structure and the IEEE XAI framework by aligning (1) the three identified aspects of explainability, (2) the modeling stages, and (3) methods for explanation assessment. Building on this information collection, MQG4AI's generic and customizable lifecycle blueprint, as well as flexible information blocks ensure that the chosen explanation method(s) align with the intended goals, and provide the necessary clarity and actionability for the given context. For instance, the factors that define explanation methods within the proposed leaf-QG *Configuration* structure, are automatically linked with other dependencies, including lifecycle phases and design decisions.

Leaf-QG Contribution *QG_FidelityRobustnessScore_(SHAPLIME)* functions as a possible risk control for *Unfaithful Explanations*. Its primary objective is to contribute design knowledge to MQG4DK, while illustrating the structure of a single and purely technical leaf-QG contribution. Overall, the proposed systematic information processing layout in form of different and related information layers aims to ensure quality by design. Highlighting information integration with relevant contextual information, and QGs to address interdependencies, leaf-QGs are envisioned to act as gatekeepers, ensuring consistency and reliability in design decision-making. From a conceptual viewpoint, the proposed leaf-QG structure needs to be evaluated further for applicability, especially in the context of MQG4A, including if the intended purpose towards compliance is fulfilled.

MQG4A Scenarios MQG4A reflects concrete use cases, which requires a different approach to MQG4AI's building blocks than with MQG4DK. Depending on the use case, a combination of explanations may be used, leading to multiple MQG4A template versions if tested for the same purpose, and/or multiple leaf-QG *Configuration* within a single MQG4A-version to address different explanation purposes that apply simultaneously. They each correspond to concrete results for collection-QG *Evaluation* and implemented qualitative or quantitative metrics, as well as information transformation as organized within collection-QG *User Interpretation*. Therefore, unlike leaf-QG *Configuration* (of which multiple may exist), the collection-QGs may encompass various approaches used to assess/translate any XAI method's results within MQG4A. Method assessment is repeated for all existing XAI methods within a single project that each relate to their respective *leaf-QG Configuration*. This flexibility is critical during development and when evaluating system updates, because some explanation methods, while effective in certain contexts, may not meet the specific needs of a given use case [ZJ21b]. Mapping these needs to the appropriate evaluation metrics is necessary to ensure that explanations continuously meet their intended goals.

MQG4AI & *Transparency* Finally, while the *IEEE Guide for an Architectural Framework for Explainable Artificial Intelligence* (XAI) [Art24] focuses on explanations, *Transparency* is considered across multiple dimensions. These include *simulatability*, which analyzes how model behavior shifts with varying data [Art24, 23]; *decomposability*, which emphasizes understanding the model's individual components; and *algorithmic transparency*, which clarifies how specific input combinations yield particular outputs [Art24, 24]. While these aspects enhance interpretability and trust, they are not directly tied to explainability. MQG4AI is designed to contribute relevant information to these transparency-related factors by incorporating structured lifecycle planning, versioning, and interdependent information processing. The, in this chapter, introduced *explanation* lifecycle stage focuses on reliable design decision-making, ultimately supporting *Transparency* through guided information management (IM).

In summary, this chapter demonstrates a generic *explanation* lifecycle stage, advancing the design of the living MQG4AI lifecycle planning blueprint. Additionally, it introduces the process of contributing leaf-QGs within MQG4DK. To further assess the proposed workflow's suitability for developing the RAI design knowledge base, other possible leaf-QG contribution scenarios require further testing. Therefore, the next chapter 7 offers and alternative perspective and addresses a contribution to MQG4DK that comprises a compilation of leaf-QGs centered around a fictional use case.

7 MQG4DK Design – Leaf-QG Compilation towards Reliable Performance Evaluation Metrics based on a Fictional Use Case

After introducing the leaf-QG information processing template for a concrete example contribution to MQG4DK, and illustrating how to design the *explanation* lifecycle stage in chapter 6, this chapter explores a more complex MQG4DK contribution in form of a leaf-QG compilation towards reliable performance evaluation centered around a fictional use case for ECG multi-label classification in emergency medicine (EM), as detailed in section 5.2.1. This chapter is based on [EM24] and [EB23], where we introduce MQG4AI, while exploring reliable performance evaluation, and the former is centered around multi-label ECG classification. More information on our experiments can be found on Zenodo[1], and the QG-compilation within MQG4AI can be viewed on GitHub[2]. We aim to provide design knowledge to address risks assigned to the TAI requirement *Technical Robustness and Safety-Requirement* focusing on *Accuracy*, as introduced in sections 3.2 and 3.3.5.2. Next, we simulate MQG4A for model configuration using *pre-*, *intra-*, *post-selection* MQG4A template versions in chapter 8. The iterative method for reliable design decision-making is derived from our experiments towards reliable performance metrics, which is introduced as a contribution to MQG4DK in this chapter.

First, we illustrate components that belong to a reliable metrics selection through information linking along the MQG4AI lifecycle blueprint. This approach is derived from a retrospective evaluation of our experiments [EM24]. Therefore, section 7.1 further outlines the related RM and system information blocks within MQG4AI. Section 7.2 explains our proposed risk-mitigating leaf-QG compilation, highlighting domain-specific considerations, as well as related design decisions along the AI lifecycle. Concluding, section 7.3 evaluates our proposition.

[1] https://zenodo.org/records/10931084

[2] https://github.com/miriamelia/MQG4AI/blob/main/README.md

7.1. Supplementary RAI Knowledge

Embedding the proposed setup within a fictional use case intends to enable the illustration of MQG4DK interaction, under the inclusion of MQG4AI's supplementary information blocks RM, and related system information. As a result, the, in this section explored information collection shapes the approach to RAI design knowledge extraction, as demonstrated in section 7.2 next. Therefore, the resulting leaf-QG compilation is based on a fictional use case situated in EM, as introduced in section 5.2.1, highlighting the domain-specific label structure, which is directly linked to evaluation metrics interpretation.

7.1.1. Risk Management

We identify two related risks, *Unreliable Performance Metrics* is assigned to *Accuracy* under *Technical Robustness and Safety* within the Assessment List for Trustworthy AI (ALTAI) [Hig20]. In addition, due to the required domain (and data) knowledge for reliable performance metrics, we broach the risk *Lack of a Domain Experts and Collaboration Mechanisms*, assigned to *Diversity, Non-discrimination and Fairness*.[3] Other relevant risks emphasizing data distributions, which impact metrics' suitability, for instance, are not the focus of our example, and we are working with a predefined open-source dataset (Physionet [RM21a]).

Horizontal Standardization Generally, technical knowledge on metrics calculation and interdependent design decisions is important to choose metrics that align with the respective structural setting. ISO/IEC TS 4213 on the *Assessment of machine learning classification performance* [ISO22b] provides a solid compilation of relevant performance metrics, including information on how to embed the selection process with the AI lifecycle. However, currently, there is no standardized approach for applying performance metrics across the diverse range of medical use cases. This lack of standardization poses a risk of oversimplification and neglecting critical contextual factors [Tab23, 6]. Each medical AI project relies on a unique combination of data types, tuning objectives, and dataset distributions, all of which influence the interpretation of selected metrics [TJ21]. As a result, domain knowledge is a crucial component in identifying a fitting performance evaluation metrics compilation.

Vertical Standardization Currently, various performance metrics are used to evaluate AI models in ECG interpretation, and inconsistencies in metrics application exist [BA20, 881]. Nevertheless, standard metrics, centered around the F1-Score seem to crystallize [MS23, 4] [GA24a, 1643] [EM24, 5], which aligns with

[3] https://github.com/miriamelia/MQG4AI/tree/main/MQG4DesignKnowledge/3_RiskManagement/AI_Risks

the American National Standards Institute (ANSI)/Association for the Advancement of Medical Instrumentation (AAMI) EC57 standard, which "[...] is the [Food and Drug Administration] FDA-recognized consensus standard and provides detailed instructions for measuring beat and rhythm detection/classification sensitivity (Se), and positive predictive value (PPV)" [BA20, 884].[4] Generally, standardization of metrics with respect to concrete application scenarios is a critical factor in enabling comparability of different algorithms, as well as bench-marking against other approaches, and state-of-the-art methods. For instance, aiming to establish an ECG-related metrics compilation, complementing the F1-Score with insights on the negative label can be beneficial. This approach would follow traditional ECG performance evaluation, which considers specificity, i.e. sensitivity for the negative label [MK21a].[5]

7.1.1.1. Technical Robustness & Safety: Unreliable Performance Evaluation

From a structural viewpoint, ECG classification comprises different formats. Use cases include binary, multi-class (one of three or more possible labels), as well as multi-label (multiple labels that may coexist) classification scenarios, shaping the precise calculation of metrics. This section provides a selection of risk sources that impact the suitability of performance evaluation metrics, focusing on our proposed risk mitigating QG contribution, as outlined in section 7.2. Highlighting our fictional use case, we concentrate on multi-label ECG classification in EM, which impacts metrics interpretation, as well as the underlying label structure. The following summarizes a collection of identified metrics-related design decisions that impact the reliable choice of performance metrics and pose risks if not implemented correctly.

Multi-Label Metrics For performance evaluation, the multi-label case is often transformed into multiple independent binary classification problems, where each label is assessed as present vs. not present. An averaging method is then applied to derive overall classification metrics. Multi-label metrics such as the *exact match ratio* (the proportion of instances where all predicted labels exactly match the true labels), or *hamming loss* (the proportion of incorrect labels per

[4] The F1-Score is the harmonic mean of PPV, or Precision (i.e. correctly as true predicted labels out of all true predictions) and Se, or Recall (i.e. the amount of true labels, the model predicted correctly out of all existing true labels).

[5] As an example to illustrate diverging metrics, the myocardial infarction multi-class classification model evaluation [GS22], as introduced in section 5.2.1 does not include the F1-Score or related metrics, and is evaluated with the C-statistic (ROC AUC) to assess the model's discrimination capabilities. However, ROC AUC tends to be too optimistic for highly imbalanced data sets [EM24, 5], which equally applies to their data, as outlined in the supplementary material [GS22], thus demonstrating inconsistencies and possible pitfalls in metrics selection. Additionally, they include other calibration metrics focusing on the model's overall suitability, as we will broach in section 7.2.

sample) [ISO22b, 14], as well as averaging methods [ISO22b, 12] are equally introduced in ISO/IEC TS 4213. Multi-label specific inaccuracies may arise, since this binary metrics transformation overlooks label correlations [ZZ14, 1820].[6] Aiming to consider label correlations through domain knowledge, comprises additional design decisions that may be centered around post-processing thresholding methods, as introduced in more detail for our experiments as a risk mitigating measure in the next section 7.2.

Imbalanced Data Imbalanced data is common in medical applications, requiring stakeholders to carefully select and interpret metrics in alignment with the use case at hand. When data is imbalanced, focusing on labels, the positive and negative classes are unequally distributed, while usually the, for medicine interesting true class, i.e. disease data, is heavily underrepresented compared with negative, healthy data samples. For instance, unlike sensitivity and specificity, *positive and negative predictive value* (PPV/NPV) are influenced by the proportion of disease and healthy cases in the test set [TJ21, 5].[7] Also, accuracy alone is insufficient in imbalanced scenarios, as it can be dominated by the majority class (usually reflecting healthy samples). Therefore, alternative performance metrics and related learning and resampling strategies should be considered [TJ21, 13]. ISO/IEC TS 4213 equally addresses performance evaluation-related design decisions highlighting bias, data representativeness, preprocessing, and the model training strategy, among others [ISO22b, 5].

Metrics Selection From the recipient's perspective, a comprehensive understanding of the information conveyed is crucial for interpreting ML performance metrics, particularly in medicine, though it is not always guaranteed [HSA22, 1]. For example, the ROC AUC is a common metric used for classification tasks, including for popular benchmarks like the STOIC challenge[8] for COVID-19 lung classification [BL24]. Their selection of metrics is based on [RA21], which states that both ROC AUC and its counterpart, PR AUC, reflect data imbalance [RA21, 48]. However, there is an ongoing debate on whether ROC AUC accurately reflects the performance on imbalanced datasets, a common scenario in medicine [DJ06, ST15]. A profound study on metrics, including effects of different classifiers, as well as training methods like *cross-validation* is discussed in [TJ21][9], where the authors analyze and explain classification and ranking based metrics with respect to healthcare-specific requirements, from a mathematical point of

[6] Alternative multi-label strategies account for pairwise or multiple label correlations with increasing computational complexity but are not the focus of this analysis [ZZ14, 1820].

[7] PPV is equivalent to precision in binary classification [TJ21, 6].

[8] `https://stoic2021.grand-challenge.org/`

[9] Referring to the complex interdependencies along the AI lifecycle, careful consideration is essential when designing training, validation, and test datasets, as they must be independent to ensure a bias-reduced evaluation and possibly stratified to address class imbalance. These design decisions impact model performance and a successful implementation is evaluated against the selected metrics. For a reliable interpretation, therefore, other design decisions need to be accurate and interdependencies made transparent.

view. However, PR AUC is not mentioned, and ROC AUC is not explicitly discussed for imbalanced data sets, illustrating prevailing inconsistencies. Further, within our experiments, ROC AUC performed very optimistic for rare labels and produced misleading results, compared to PR AUC. In summary, inconsistencies arise in the use and application of these metrics across papers and benchmarking tools [RA20, SN20], highlighting the need for standardization. These efforts should result in a comprehensive set of (complementary) metrics, since no single metric captures all aspects of a model's performance [HSA22, 1]. Moreover, integrating technical and domain knowledge enables a more nuanced real world embedding, supporting risk mitigation, improved result monitoring, and informed follow-up actions. Finally, the question arises, how much effort is placed on metrics selection and comprehension in practice. Given the importance of metrics selection in medical AI, further research into the most appropriate evaluation methods is necessary to ensure reliable performance assessment tailored to concrete use cases. ISO/IEC TS 4213 offers a solid starting point to guide metrics selection of classification scenarios in general.

Misclassification Cost Another common consideration in healthcare is *Cost-Sensitivity*, where the cost of misclassification depends on the incorrectly predicted label [TJ21, 4], which is measured in the real world. Therefore, threshold-dependent metrics, that rely on the confusion matrix, should be tuned and tested for multiple thresholds across different data mixtures [HSA22, 2] to ensure an optimized setup. These metrics are calculated after applying a threshold to assess the model's predictive confidence, which can be challenging, particularly in multi-class or multi-label problems, which are common in medicine [TJ21, 6]. Threshold optimization is a standard tuning practice, which is expected to perform better in real world scenarios under the inclusion of domain knowledge when defining costs, compared to other methods like ROC curve analysis [TJ21, 8], for instance. This is realized for multi-label ECG classification based on a *benefit matrix*, as proposed in [LY21], which we adapt to our fictional use case [EM24], and related QGs are outlined in the next section 7.2. This complexity when aiming to accurately measure the intelligent system's performance in the real world creates additional requirements for selecting appropriate metrics, especially in use cases like multi-label ECG classification, where misclassifying different arrhythmias or diagnoses (HARHD) carries varying degrees of severity [LY21, 2]. Finally, developing custom, use-case-specific evaluation metrics that incorporate domain knowledge is a promising approach [PJ21b], as demonstrated in [RM21a], where a scoring function was created for multi-label ECG classification.

Concluding, AI performance evaluation (in medicine) would benefit from established standardization methods that bridge the gap to use case-specificity. In addition to profound technical AI literacy, domain knowledge contributes critical information from multiple perspectives. The true success of any application can only be measured in the real world, and the inclusion of domain experts along the AI lifecycle early on is crucial for RAI.

7.1.1.2. Diversity, Non-Discrimination & Fairness: Lack of Domain Experts & Collaboration Mechanisms

The development of use-case-adapted metrics and the accurate interpretation of evaluation metrics in alignment with real world settings, necessitate the involvement of domain experts and the establishment of effective communication channels across the AI lifecycle. The absence of such interdisciplinary collaboration introduces risks related to the validity of performance evaluation metrics and other critical lifecycle design decisions. The following introduces a selection of related key points, as guiding directions, where medical domain knowledge and performance evaluation metrics need to be aligned.

Medical Context A comprehensive understanding of the clinical setting, highlighting the intelligent system's objectives and tasks, is essential for metrics selection. Therefore, medical AI in general should be supported by an interdisciplinary team, formed through the identification of relevant disciplines, and the continuous recognition of interdependencies throughout the AI lifecycle to enable a long-term monitoring strategy that mitigates long-term performance degradation, which is measured based on the selected performance metrics compilation. The absence of these key consideration introduces several risks, such as misalignment with clinical needs, a biased or incomplete evaluation, as well as increased patient risks, among others.

Metrics Interpretation Evaluating clinical efficacy is inherently complex [KC19, 3], and in addition to medical knowledge, as a general principle, each selected metric should be supplemented with additional material, relating to the data distribution from different angles. This supports precisely documenting metrics' interpretation within the real world medical context, while shedding light on the underlying and predicted data distributions, enhancing metrics interpretation and suitability evaluation by design. Domain-embedded evaluation approaches are considered the most effective and should be the standard in medicine, as they assess an intelligent application's real world performance and impact [KC19, 3]. In addition to linking a domain-independent metric's output with domain-specific real world characteristics, "[m]any fields of biomedicine have published their own guidelines on how to evaluate machine learning algorithms [...]" [TJ21, 9], as with the previously introduced Physionet scoring metric [RM21a].

Data Interpretation A comprehensive understanding of the underlying data is essential, as it serves as the critical link to real world clinical settings. However, clinical data is often distributed, heterogeneous, and high-dimensional, requiring the integration of multiple sources based on medical reasoning to generate meaningful input for ML models [MH21, 120]. This fusion process can be highly complex and resource-intensive. Consequently, "[t]he development of quality recommendations and standards for training data sets has to be a community-driven effort of many diverse stakeholders" [MH21, 120], as high-quality datasets are a crucial determinant of model performance.

In summary, this necessary additional knowledge to implement a risk mitigated performance evaluation metrics compilation requires use case-specific information on the AI system. Therefore, the next section summarizes system information for our fictional *Custodian* use case in EM, while highlighting information on the application and stakeholder information modules, as introduced in section 3.3.3.1.

7.1.2. System Information

Aiming to link relevant domain knowledge with metrics selection, specific to our fictional use case, we append information on the AI application to our MQG4DK contribution scenario. For our exemplified metrics selection snapshot, we define a fictional AI with a *Custodian* functionality in EM to assist non-specialized medical personnel with ECG interpretation, as detailed in section 5.2.1. This includes the provision of domain knowledge on ECGs, accessible to all stakeholders through shared MQG4A lifecycle planning templates, once this design knowledge is applied. Additionally, further addressing the previously introduced risks, we explore required stakeholder roles who participate in metrics selection, highlighting the inclusion of a consulting medical domain expert.

7.1.2.1. Application

The *application* module within the system information block comprises foundational domain knowledge on ECGs, the human heart, as well as heart diseases[10] and we describe the envisioned *Custodian* AI with an alarming guard functionality based on the concepts as identified in the AIRO [GD22], viewable on GitHub[11].

Use Case The intelligent medical product at the core of our research is designed as a CDSS for settings where patients at risk may be present, but the attending personnel are not well-trained in ECG diagnostics. For instance, this includes ambulances, emergency scenarios, or general practitioners' offices, while we focus on EM for our design. Consequently, the clinical follow-up step after the model's

[10] `https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/1_System/Application/DomainKnowledge/Electrocardiogram_(BasicKnowledge).md`

[11] `https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/1_System/Application/example_ECGAlarmingGuardFunctionality_(EmergencyMedicine).md`

prediction plays a crucial role in shaping other AI system-related design decisions that directly impact the patient. Based on the medical decision supported through physician-AI collaboration, the patient is either treated immediately in urgent cases, sent home if deemed healthy, or referred to a hospital for further evaluation.

Domain Knowledge Domain knowledge is fundamental for trustworthy design decision-making and must be continuously integrated throughout the AI lifecycle. For instance, in our use case, balancing the importance of precision (*how many of the, as positive identified labels were correct?*) versus recall (*how many of the existing positive labels were identified?*) depends on the specific predicted label, which reflects identified HARHD(s). This dependency raises key questions about metrics selection (objective), training strategy (metric-to-monitor), optimization (thresholding) and evaluation (averaged multi-label metrics and per-label performance): *Should the system prioritize sensitivity (i.e. recall), accepting a higher false alarm rate to avoid missing critical cases, or should false positives be minimized at the risk of missing some conditions?*

7.1.2.2. Stakeholder

As became evident throughout the previous sections, especially in the medical domain, the involvement of domain experts is essential to ensure the accurate transfer of complex domain knowledge required for the selection of reliable performance evaluation metrics. Within our experiments, an anesthetist[12] provided the necessary expertise to support us, the developers[13] in making informed design decisions. Additionally, a RAI person[14] is included within our proposition to consider the responsible implementation of performance evaluation metrics in a risk mitigated manner. Furthermore, the output of the intelligent system, in conjunction with medical personnel who hold the final decision-making authority, directly affects patient[15] outcomes. Thus, for our proposed snapshot, we integrate *active* (developer, RAI person), *consulting* (domain expert), and a *passive* (patient) stakeholder roles with the MQG4DK contribution within the system-related information block. Additional relevant stakeholders, such as organizational personnel within healthcare companies or auditors responsible for regulatory compliance, equally play an important role. However, their consideration falls beyond the scope of our contribution. For a more comprehensive perspective

[12] `https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/1_System/Stakeholder/2_Consulting/DomainExpert_(ConsultingStakeholder).md`

[13] `https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/1_System/Stakeholder/1_Active/Developer_(ActiveStakeholder).md`

[14] `https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/1_System/Stakeholder/1_Active/ResponsibleAIPerson_(ActiveStakeholder).md`

[15] `https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/1_System/Stakeholder/3_Passive/User_(PassiveStakeholder).md`

on stakeholder inclusion, refer to section 3.3.2.1, where we introduce MQG4AI's fundamental design principles, and section 3.3.3.5, where we explore existing roles in alignment with the AI Act.

After introducing ECG multi-label classification performance evaluation related risks and AI system information centered around our fictional *Custodian* use case, the next section outlines the proposed QG compilation along the AI lifecycle towards reliable metrics.

7.2. Proposed Leaf-QG Compilation along the AI Lifecycle

After introducing performance evaluation metrics related risks, and the fictional *Custodian* use case in EM, filling MQG4AI's supplementary information blocks, this section highlights our proposed risk control in form of a leaf-QG compilation. It comprises a combination of related design decisions for reliable performance evaluation metrics selection. Focusing on multi-label ECG classification in an emergency setting, the MQG4DK contribution is intended to provide generalizable guidelines for (multi-label) classification metrics during model development, as well as illustrate how to include medical domain knowledge along the AI lifecycle. A robust evaluation of AI performance requires selecting multiple metrics to ensure a comprehensive assessment that accommodates different stakeholder perspectives. Since no single metric can capture all desired capabilities [KC19, 3], the first step in designing an appropriate metric compilation is to analyze the interplay of all relevant use case requirements, ideally before implementation. This includes considering data characteristics, tuning objectives, the contextual meaning of each metric, and potential use case-specific limitations that may necessitate creative workarounds.

Continuing considerations on AI output examination in section 2.2, in the multi-label scenario, key factors include label correlations, potentially varying per-label tuning objectives, and the interpretability of averaged multi-label metrics. The following technical guidelines are included within the identified leaf-QGs that we introduce in this section:

- A comprehensive selection of ML performance metrics, including confusion-matrix-based and ranking metrics, chosen based on their use-case-specific interpretability

- A careful assessment of the ROC AUC metric, considering its limitations and applicability

- A detailed analysis of per label/class probabilities, supplemented by additional material to enable a thorough understanding of metric interactions and their dependence on data distribution

- The application of macro- or weighted-averaged multi-label metrics to ensure a more balanced evaluation of the model's performance

- Considerations for a well-defined strategy for metric optimization and monitoring throughout the system's lifecycle

- An in-depth analysis of use-case-specific challenges, particularly regarding label correlations in multi-label classification, along with potential solutions

Our proposed QG compilation, as introduced in this section, aims for gatekeeping by design, meaning that if the identified QGs are implemented and integrated with the AI lifecycle, the classification metrics selection is reliable. Their concrete realization is more use case-specific. To derive the QG compilation, we begin with the core design decision, i.e. reliable performance evaluation metrics and derive separate, related design decisions such as a use case-adapted label interpretation, or the inclusion of additional material for metrics interpretation with respect to data distributions, based on our experiments [EM24] [EM25c].

7.2.1. Experimental Setup

Before diving deeper into QG lifecycle design, we briefly introduce our experimental setup. More comprehensive information can be viewed in detail on Zenodo [EM25c]. The fictional *Custodian* use case, including our approach to design a fitting label structure, is introduced in section 5.2.1. Overall, we employed a basic training setup for this stage, as our primary focus is on process extraction and the analysis of interrelated tendencies in AI evaluation metrics, rather than performance optimization.

Training The model architecture follows the design proposed in [XX20], combining bidirectional LSTMs and CNNs. During training, we incorporated early stopping and limited the maximum training duration to 50 epochs, with a batch size of 32. Additional hyperparameters were adapted from [LY21], including an initial learning rate of 0.001 with Exponential Decay, the Adam Optimizer, and Binary CE. To ensure a balanced distribution of labels across all data subsets, the experiments were conducted using a proportional, stratified split of 70% training, 10% validation, and 20% test data. This approach guarantees the representation of rare labels in each subset, while preserving the underlying data distribution, which impacts model performance. Specifically, the relationship between the training data and the unseen real world data encountered post-deployment may impact generalizability. However, given our fictional setting, an in-depth evaluation of these aspects lies beyond our current scope.

Data Our experiments utilize publicly available data from the PhysioNet Challenge 2021 [RM21a], where participants aimed to develop models capable of identifying clinical diagnoses from ECG recordings. Each sample may contain multiple, interrelated HARHDs, which, from a technical perspective, corresponds to multi-label classification. For this study, we exclusively focus on 12-lead ECGs to ensure consistency in the input data. To incorporate the medically relevant label correlations, our domain-embedded procedure is based on *Category Imbalance and Cost-Sensitive Thresholding* (CIST) [LY21]. This approach enhances model evaluation by aligning decision thresholds with the clinical significance of different diagnoses, ensuring a more reliable assessment of model performance in a real world medical context. We derive our use case-specific label structure by mapping the PhysioNet dataset to our classification framework using SNOMED CT codes[16]. Overall, Physionet data is notably diverse, comprising "[...] ten databases from [...] several countries across three continents" [RM22, 3], including the US, Europe, and China. This diversity offers valuable insights into the behavior of performance metrics, as it reflects the variability in clinical data across different geographic regions and healthcare systems. Further, it supports understanding how performance metrics may behave in real world settings.

Results We identified several key tendencies influencing design decisions related to performance evaluation metrics. A crucial aspect is the use case-adapted metrics selection, ensuring that chosen metrics align with the real world application (1). Additionally, we analyzed the expressiveness of ROC AUC vs. PR AUC, particularly in the context of imbalanced medical datasets, where PR AUC may better reflect model performance (2). To address the challenges of multi-label classification, we explored different multi-label averaging strategies, balancing metric interpretation across labels of varying prevalence (3). Further, we considered label structures based on label support, evaluating how each label's frequency in the dataset and its relative representation influence model evaluation (4). Another key factor was selecting the most appropriate metric for training optimization on the validation data set, influencing early stopping and hyperparameter tuning (5). Lastly, we examined post-processing thresholding methods, comparing a fixed threshold of 0.5, CIST [LY21], and ROC and PR curve-based thresholds to better understand the behavior of rank-based metrics (6).

The following sections are intended as a template to guide the systematic definition of reliable performance evaluation metrics along the AI lifecycle (MQG4A), resulting in a contribution to MQG4DK. This structured approach aims to align performance evaluation with both, clinical requirements and the technical constraints of DL. We begin with relevant considerations on *Data Utilization*.

[16] `https://www.snomed.org/`

7.2.2. Data – Utilization

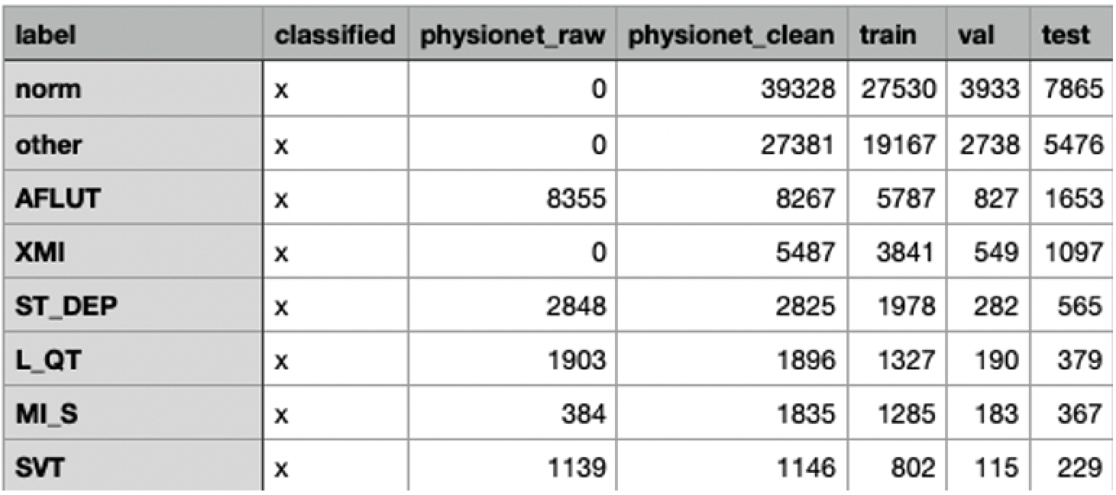
Our study begins with an existing dataset, namely PhysioNet [RM21a], meaning that certain design decisions preceding data utilization, as well as some aspects during its use, were not our focus. For example, we did not implement optimized data cleaning or a comprehensive data preparation stage, specifically tailored to model training. Instead, our approach to performance metrics selection during data utilization concentrates on data transformation, particularly the development of a use case-adapted multi-label structure, resulting in *QG_LabelStructure_(MultiLabelClassificationPreprocessing)*[17]. The QG's realization is shaped by medical domain knowledge and the label support within the dataset, ensuring that the evaluation process aligns with the real world clinical context, as introduced in section 5.2.1. Labels are particularly relevant to performance metrics selection since they comprise the model output that is being evaluated.

7.2.2.1. Medical Label Structure

In our multi-label setting, labels represent all possible HARHDs that can be detected in an ECG recording, allowing for their potential coexistence. As a foundation for our use case-adapted and domain-embedded label compilation, we relied on the PTB-XL mapping [SN20], refining it by excluding clinically irrelevant artifacts. Notably, not all labels necessarily equate to a diagnosis in the context of EM, where immediate clinical decisions are required.

To align label structuring with real world decision-making, particularly in an emergency setting, we introduced two novel container labels: *other* and *norm*. The *norm* label represents a healthy patient, encompassing all unproblematic ECG patterns, such as sinus rhythm variants and physiological electrical cardiac axis. It is mutually exclusive and cannot coexist with any other label. Conversely, the *other* category includes pathological ECG findings that do not directly indicate a specific therapy without further diagnostic evaluation. If an ECG is classified under the *other* category, the patient is expected to be referred for hospitalization. *Other* can coexist with possibly, additionally identified pathologies. To further refine classification to our fictional use case in EM, we introduced an additional container label for ischemias, grouping conditions that require similar follow-up treatment (excluding ST depression and elevation). The resulting proposed label structure consists of 13 labels, and designed to enhance interpretability and facilitate more clinically meaningful decision-making in emergency scenarios. However, for our concrete experiments, not all medically reasonable labels are sufficiently represented in the open-source data, leading to further adjustments.

[17] `https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/2_Lifecycle/1_Data/2_Utilization/2_Preprocessing/2_Transformation/QG_LabelStructure_(MultiLabelClassificationPreprocessing).md`



label	classified	physionet_raw	physionet_clean	train	val	test
norm	x	0	39328	27530	3933	7865
other	x	0	27381	19167	2738	5476
AFLUT	x	8355	8267	5787	827	1653
XMI	x	0	5487	3841	549	1097
ST_DEP	x	2848	2825	1978	282	565
L_QT	x	1903	1896	1327	190	379
MI_S	x	384	1835	1285	183	367
SVT	x	1139	1146	802	115	229

Figure 7.1.: Our experimental label count for the final 8 labels, as derived from Physionet open-source data [RM21a].

7.2.2.2. Label Distribution

Considering the (dynamic) data distribution throughout the entire lifecycle is essential for accurate model evaluation and comparison. Continuous (re-) assessment of data suitability is necessary to ensure that the dataset remains aligned with the intended purpose of the intelligent system. For instance, incorporating medical prevalence rates can offer valuable insights into what constitutes a realistic label distribution, improving both, the robustness and generalizability of the model. The medical prevalence of the respective label for a specific subgroup can support applicability assessment and monitoring, as in [RK23], where the authors evaluate data drift in clinical sepsis prediction, or [SP09] for prevalence assessment of the long QT syndrome among Caucasian infants, for instance. Additionally, a data distribution monitoring strategy can help mitigate risks associated with biased model behavior, as highlighted in ISO 24027, which addresses bias in AI systems and aided decision-making [ISO21a].

Experimental Label Structure For our initial experiments, we opted to itemize the container labels, resulting in a total of 21 labels. The mapping of our label structure to the Physionet data via their SNOMED CT code introduced some incompleteness, as not all clinically relevant labels for our use case were sufficiently represented in the data. Notably, Pericarditis was entirely absent from the dataset. Our first model was trained on 20 labels. However, due to low label support, some labels could not be meaningfully evaluated, as their per-label metrics were not reliably calculated (metric value = 0). Consequently, we excluded these underrepresented labels from further analysis. This decision was grounded in ensuring that each label had sufficient representation to enable reliable evaluation. Per class metrics could not be calculated for AFIB_T[18], several fine-grained ischemia, AV2, AV3, ST_ELE, and VT, and they were excluded. Ultimately, we proceeded with 8 labels (AFLUT, L_QT, MI_S and XMI (both include all ischemia for a higher label count), ST_DEP, SVT, *norm* and *other*), selecting those with at least ≈ 1100 samples, which remains a relatively low threshold. The subsequent processes and results presented in this study are based on these 8 labels, as depicted in Figure 7.1.

Before introducing the proposed *Development* stage related leaf-QG compilation, we first highlight design decision-making relevant information. Concretely, we propose to summarize open-source data sources as RAI design knowledge contributions. Further, we propose an iterative design decision-making method categorized into pre-selection, intra-selection, and post-selection steps [EM24], which equally shapes MQG4A template versions in the next chapter 8.

7.2.3. Development – Design Decision-Making

Due to the inherent dynamics of AI, the current state of AI design knowledge, and the use case-specific nature of AI applications, design decision-making is an iterative and experimental process, particularly during the development phase. To address this, we incorporate a dedicated design decision-making section within the development stage of the MQG4AI lifecycle blueprint.[19] This additional information collection is intended to enhance AI literacy and compliance, while also structuring methodological knowledge in a systematic manner.

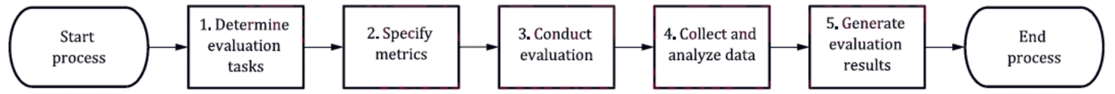

Start process
1. Determine evaluation tasks
2. Specify metrics
3. Conduct evaluation
4. Collect and analyze data
5. Generate evaluation results
End process

Figure 7.2.: Process steps for performance assessment, as introduced in ISO/IEC TS 4213 [ISO22b, 4].

ISO/IEC TS 4213 on the *Assessment of machine learning classification performance* introduces a "[g]eneralized process for machine learning classification performance assessment" [ISO22b, 4], as depicted in Figure 7.2. The process comprises five high-level stages. After defining the classification task (1), the metrics are identified (2). Next, the evaluation plan is conceptualized and implemented, including information on the data, software and hardware (3). After model training, the raw model outputs, generated for each sample, are collected (4). Finally, the evaluation in conducted based on the specified metrics and possibly, additional

[18] For our use case, the distinction into two types of AFIB that require a different treatment would have been important but was not made in the original data.

[19] `https://github.com/miriamelia/MQG4AI/tree/main/MQG4DesignKnowledge/2_Lifecycle/2_Development/0_DesignDecisionMaking`

relevant information (5). [ISO22b, 4]. Complementing this workflow proposition, we follow an information-oriented approach within our proposed generalizable three-fold design decision-making method. We propose to organize relevant information on a concrete design decision according to pre-, intra-, and post-selection steps.[20] Thus, focusing on (2), we identify relevant information for reliable metrics selection, which includes an iterative evaluation of (3) (4), and (5). Further, (1), as well as partly information of (3) is acquired beforehand. In addition, based on our experiments, we emphasize the application of open-source data, which is equally relevant for scientific publications and knowledge dissemination. Especially, in case the data is privately owned and information thereof preferably kept private, open-source data enables public knowledge sharing.

7.2.4. Development – Evaluation

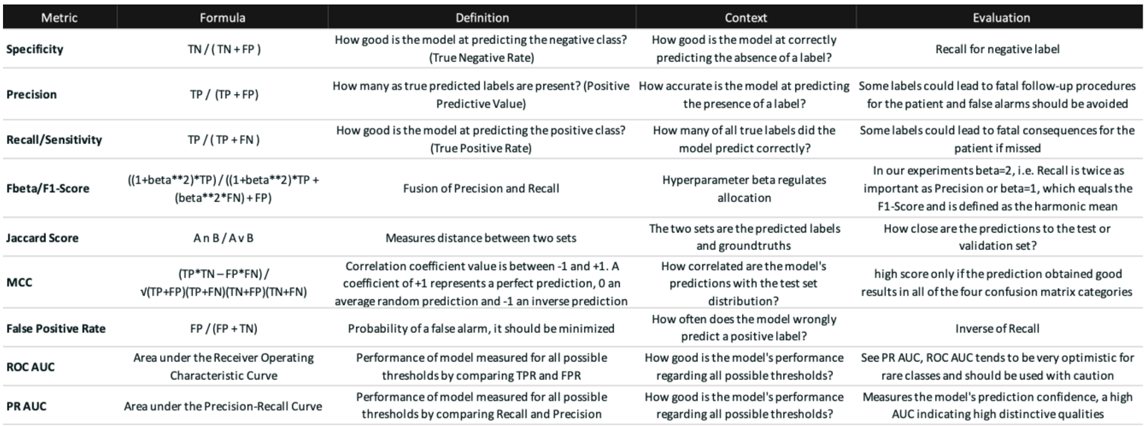
The lifecycle development *Evaluation* stage is at the center of our proposed MQG4DK contribution. The following sections introduce three main design decisions that we derived based on our experiments. Each decision comprises a larger, interconnected choice that both influences and is influenced by the others. This interdependence underscores the complexity of decision-making, where a change in one aspect can have ripple effects throughout the entire process. We begin with the performance metrics compilation, highlighting technical knowledge and domain-specific considerations for metrics selection. In our multi-label case, a separate leaf-QG to define training objectives tailored to each label is appended. This is relevant for metrics interpretation. Finally, as previously mentioned, performance evaluation metrics should be accompanied by additional material, resulting in a third collection-QG comprising individual leaf-QGs.

7.2.4.1. Performance Metrics Compilation

Overall, distinct metrics measure different qualities, which necessitates a combination of metrics for a comprehensive model evaluation, as illustrated in Figure 7.3 for our fictional use case [EM24, 7], highlighting metrics such as Sensitivity and Specificity [MK21a, 444], which are utilized for ECG evaluation.[21] Generally, metrics proposed in the literature for AI-ECG evaluation comprise confusion matrix-related metrics, as well as ranking-based metrics [EM24, 3] [MS23,

[20] This is simulated for MQG4A in chapter 8, and the process structure may qualify as a generic guideline for adapting MQG4DK contributions to MQG4A scenarios in form of different template versions to organize concrete results.

[21] `https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/2_Lifecycle/2_Development/2_Model_Evaluation/PerformanceMetrics/QG_PerformanceMetricsCompilation_(MultiLabelClassification).md`



Metric	Formula	Definition	Context	Evaluation
Specificity	TN / (TN + FP)	How good is the model at predicting the negative class? (True Negative Rate)	How good is the model at correctly predicting the absence of a label?	Recall for negative label
Precision	TP / (TP + FP)	How many as true predicted labels are present? (Positive Predictive Value)	How accurate is the model at predicting the presence of a label?	Some labels could lead to fatal follow-up procedures for the patient and false alarms should be avoided
Recall/Sensitivity	TP / (TP + FN)	How good is the model at predicting the positive class? (True Positive Rate)	How many of all true labels did the model predict correctly?	Some labels could lead to fatal consequences for the patient if missed
Fbeta/F1-Score	((1+beta**2)*TP) / ((1+beta**2)*TP + (beta**2*FN) + FP)	Fusion of Precision and Recall	Hyperparameter beta regulates allocation	In our experiments beta=2, i.e. Recall is twice as important as Precision or beta=1, which equals the F1-Score and is defined as the harmonic mean
Jaccard Score	A n B / A v B	Measures distance between two sets	The two sets are the predicted labels and groundtruths	How close are the predictions to the test or validation set?
MCC	(TP*TN – FP*FN) / √(TP+FP)(TP+FN)(TN+FP)(TN+FN)	Correlation coefficient value is between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 an average random prediction and -1 an inverse prediction	How correlated are the model's predictions with the test set distribution?	high score only if the prediction obtained good results in all of the four confusion matrix categories
False Positive Rate	FP / (FP + TN)	Probability of a false alarm, it should be minimized	How often does the model wrongly predict a positive label?	Inverse of Recall
ROC AUC	Area under the Receiver Operating Characteristic Curve	Performance of model measured for all possible thresholds by comparing TPR and FPR	How good is the model's performance regarding all possible thresholds?	See PR AUC, ROC AUC tends to be very optimistic for rare classes and should be used with caution
PR AUC	Area under the Precision-Recall Curve	Performance of model measured for all possible thresholds by comparing Recall and Precision	How good is the model's performance regarding all possible thresholds?	Measures the model's prediction confidence, a high AUC indicating high distinctive qualities

Figure 7.3.: The metrics compilation that builds the foundation for our MQG4DK-contribution scenario, aiming for a comprehensive evaluation of model performance with respect to the clinical setting. Additional metrics, e.g. evaluating model calibration [GS22], can be appended.

3] [GA24a, 1643], highlighting PR AUC for imbalanced data, for instance, and possibly including use case-specific metrics, such as the Physionet Scoring Metric[22].

Classification Metrics The majority of existing binary classification metrics are calculated with the four values of the confusion matrix [TJ21, 5]. The model's predictions are compared to a threshold (set to 0.5 for balanced data), then categorized into either *true* or $false$ for binary classification. These results are then compared with the ground truth based on whether they belong to the *negative* or *positive* label. In the context of our experiments, we primarily focus on the interplay and performance of Specificity, Precision, Recall, F1-Score, and F-Beta Score, as their medical interpretation is well-founded, and especially Recall and Specificity are commonly used in ECG classification [MK21a]. To provide a more comprehensive overview, we also include additional metrics: the Jaccard Score (which provides insight into the distribution of intersections between predictions and ground truth), the False Positive Rate (FPR), which offers an additional perspective on false alarms, and the Matthew's Correlation Coefficient (MCC), which considers all four values of the confusion matrix. Most of the selected metrics range from 0 to 1, where a higher value indicates better performance. However, the MCC ranges from -1 to 1, with a value closer to 1 indicating better performance, and FPR should be minimized.

ROC and PR AUC As previously discussed in section 7.1.1.1, there is an on-going debate regarding the suitability of ROC AUC for heavily imbalanced datasets, and whether PR AUC should be preferred in such cases. These metrics, based on ranking perspectives, assess the real-valued function learned by the model.

[22] `https://github.com/physionetchallenges/evaluation-2021`

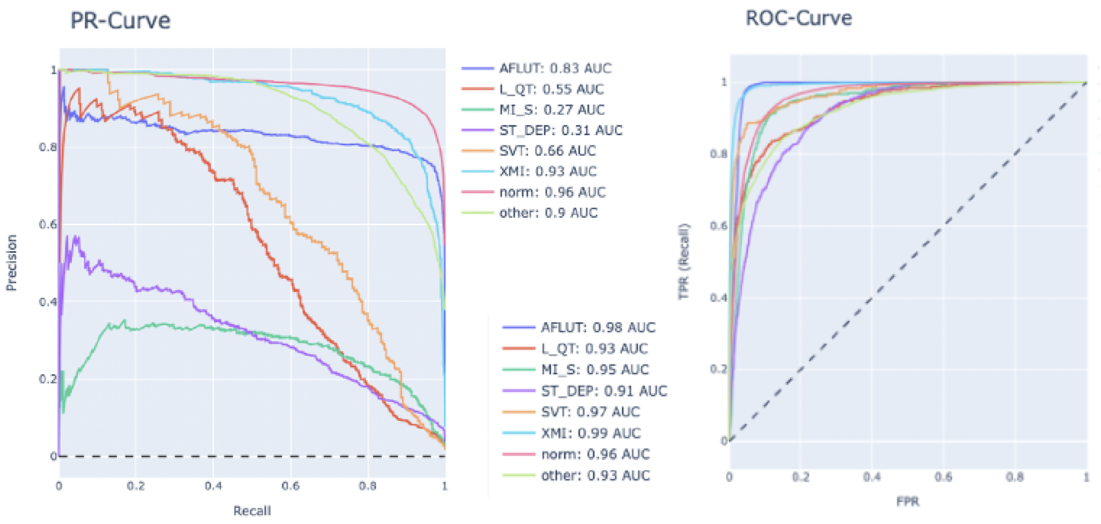


Figure 7.4.: PR AUC vs. ROC AUC performance visualized for 8 labels.

ROC AUC and PR AUC are calculated as the Area Under the Curve (AUC) for all possible thresholds concerning Recall and Precision (PR AUC), or False-Positive Rate (ROC AUC). Unlike confusion matrix-based metrics, these approaches provide a threshold-independent performance measurement and can be referenced for post-processing thresholding, as further outlined in section 7.2.5. Both metrics range from 0 to 1, with higher values indicating greater model confidence in its predictions. Our experiments confirm that, indeed, PR AUC is more informative for our multi-label use case. As shown in Figure 7.4, ROC AUC demonstrates nearly identical behavior across all labels, in contrast to the PR AUC curves, which exhibit variation, a pattern indicative and representative of label imbalance. In summary, ROC AUC tends to show overly optimistic behavior across all labels compared to PR AUC. This could potentially lead to unreliable applications in medical settings, where the actual performance of the model is worse than indicated by ROC AUC, particularly in the presence of highly imbalanced data. Therefore, for use cases that are structurally similar to our fictional *Custodian* application, ROC AUC is not deemed a suitable evaluation metric.

Alternative Angles More interesting perspectives on model evaluation and thus metrics selection that we did not consider for our exemplary MQG4DK contribution scenario exist. Further classification-related metrics are introduced in ISO/IEC TS 4213 [ISO22b], for instance highlighting statistical significance tests. For instance, the in section 5.2.1, introduced multi-class model to support diagnosis of myocardial infarction [GS22] is evaluated regarding its calibration capabilities using two metrics: the Expected Calibration Error and the Brier score [GS22, 12], measuring "[...] if the model's probability estimates reflect the ground truth empirical class frequencies [...]" [GS22, 12]. The model's calibration capabilities check how well model probability estimates align with real world frequencies, possibly resulting in a valid metrics choice for continuous model monitoring.

7.2.4.2. Additional Material

The inclusion of additional material that highlights the input and output data distribution is essential to foster a more reliable understanding of performance metrics, and should be referenced during both the planning and interpretation of metrics.[23] For instance, this supplementary material may include details about the true label and probability distribution, as well as each label's support within the data.

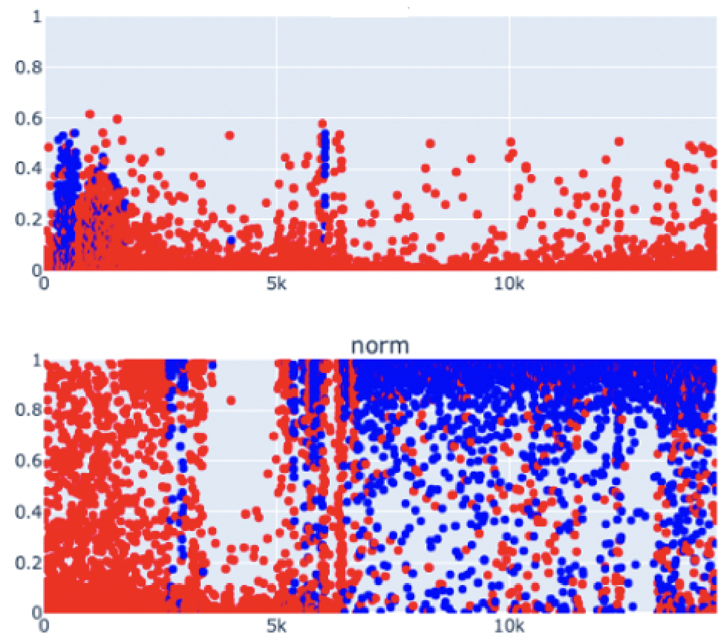


Figure 7.5.: Predicted versus groundtruth label distribution on the test set for *norm* and myocardial infarction signs (MIS).

For instance, in Figure 7.5, we visualize the ground truth versus probability distribution for the test data (with red indicating 0 and blue indicating 1) for the labels *norm* and general myocardial infarction signs (MIS). As observed, a fixed threshold of 0.5 would result in the latter label almost never being classified as *true* due to the model's low confidence. In contrast, for the majority label, *norm*, the model demonstrates the expected tendencies. To further interpret these behaviors, XAI methods could enable a deeper understanding of the model's decisions, particularly when examining the interplay between the distribution of true labels, predicted probabilities, and threshold selection. Overall, this analysis supports informed decision-making and for addresses the complexities of model performance in real world settings.

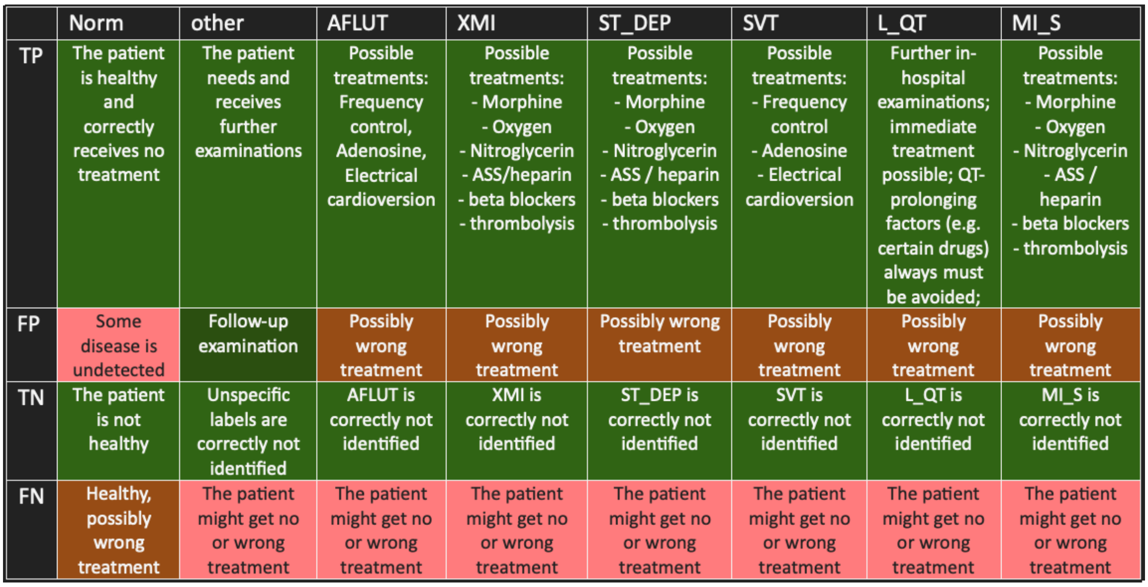
[23] `https://github.com/miriamelia/MQG4AI/tree/main/MQG4DesignKnowledge/2_Lifecycle/2_Development/2_Model_Evaluation/PerformanceMetrics/AdditionalMaterial`

7.2.4.3. Multi-Label Metrics

Building on classification-centered considerations, the multi-label case introduces additional requirements. For the initial setting, we focus on how to calculate averaged multi-label metrics from binary ones, as well as interpretation of per-label results. How to include relevant medical information on label correlations, which is essential for achieving a more contextually relevant model evaluation, is included as part of the *Optimization* stage, based on the previously introduced CIST method [LY21] in the next section 7.2.5.

Multi-label Averaging In the process of calculating overall model performance, four distinct averaging methods are typically employed: *macro* (or *weighted*), *micro*, and *samples*. The *samples* method calculates the mean performance over the entire test or validation set after evaluating each sample individually. For instance, the multi-label metric Hamming Score is derived as the samples-averaged Jaccard Score. In contrast, macro- and weighted-averaging involve calculating metrics for each label individually and then averaging them, with weighted averaging e.g. accounting for label frequency. The *micro* averaging method, on the other hand, considers the entire prediction matrix at once, which can bias the evaluation towards more prevalent labels. This approach, therefore, "[...] favors bigger classes" [SM09, 430]. [RK23, 1822], [WX16, 3] Given this, the need to analyze per-class performance in addition to the averaged metrics prevails, particularly for use cases where individual class performance is key. Additionally, some labels may prioritize Precision over Recall from a clinical perspective, or may be underrepresented in the data, making them less well-reflected by the averaged metrics that reflect global model tendencies. This highlights the importance of considering both, overall performance and per-class evaluation, especially for imbalanced datasets, to ensure that the model's behavior aligns with medical priorities and realistic use-case scenarios. As a result, macro-averaging is the preferred choice for our use case.[24] This method evaluates each label equally, ensuring that rare classes are considered [SM09, 430], and thus provides a more accurate representation of the model's performance within our fictional use case. This is especially critical in the context of ECG multi-label classification, which is marked by significant data imbalance, a common scenario to many medical use cases. As introduced, some HARHD are less frequent in the data but still medically relevant, requiring an accurate evaluation. Based on our experiments, several tendencies, related to class imbalance can be observed. When macro-averaged metrics are lower than micro-averaged metrics, this indicates that the model is less effective at identifying rare classes, but performs well on more common ones. In contrast, if macro-averaging shows higher values than micro-averaging, the model is likely to be performing more consistently across all classes, regardless of their frequency. In our case, the former tendency predominates, with only the ROC

[24] Weighted averaging suitability is assessed with respect to the selected weights, which requires further tuning, and we did not analyze this scenario for our experiments.



	Norm	other	AFLUT	XMI	ST_DEP	SVT	L_QT	MI_S
TP	The patient is healthy and correctly receives no treatment	The patient needs and receives further examinations	Possible treatments: Frequency control, Adenosine, Electrical cardioversion	Possible treatments: - Morphine - Oxygen - Nitroglycerin - ASS/heparin - beta blockers - thrombolysis	Possible treatments: - Morphine - Oxygen - Nitroglycerin - ASS / heparin - beta blockers - thrombolysis	Possible treatments: - Frequency control - Adenosine - Electrical cardioversion	Further in-hospital examinations; immediate treatment possible; QT-prolonging factors (e.g. certain drugs) always must be avoided;	Possible treatments: - Morphine - Oxygen - Nitroglycerin - ASS / heparin - beta blockers - thrombolysis
FP	Some disease is undetected	Follow-up examination	Possibly wrong treatment	Possibly wrong treatment	Possibly wrong treatment	Possibly wrong treatment	Possibly wrong treatment	Possibly wrong treatment
TN	The patient is not healthy	Unspecific labels are correctly not identified	AFLUT is correctly not identified	XMI is correctly not identified	ST_DEP is correctly not identified	SVT is correctly not identified	L_QT is correctly not identified	MI_S is correctly not identified
FN	Healthy, possibly wrong treatment	The patient might get no or wrong treatment	The patient might get no or wrong treatment	The patient might get no or wrong treatment	The patient might get no or wrong treatment	The patient might get no or wrong treatment	The patient might get no or wrong treatment	The patient might get no or wrong treatment

Figure 7.6.: Domain-embedded interpretation of the four values of the confusion matrix within our fictional use case.

AUC metric displaying high performance across both macro and micro averaging methods. This suggests that, while our model performs well in identifying the more common labels, its ability to detect rarer classes may need improvement, as reflected by the lower macro-averaged metrics.

Class Probabilities In addition to the previously introduced additional material, comparing per-label scores with the averaged multi-label metrics offers insights into the model's performance for each individual label, as well as sheds light on the expressiveness of the averaging method. Although we initially did not consider class probabilities in our experiments, the discrepancies between macro- and micro-averaged metrics prompted a retrospective analysis. By default, we calculated four metrics per class: Precision, Recall, F1-Score, and Jaccard Score. However, after conducting a deeper domain analysis, we realized that incorporating Specificity and FPR would have been valuable as well. Our experiments show that, for imbalanced multi-label classification, evaluating per-label scores provides a more accurate understanding of the model's performance, particularly since it is heavily influenced by the label's support within the data. Moreover, our use case is characterized by distinct per-label training/tuning objectives in alignment with medical knowledge.

7.2.4.4. Domain-Embedded Training Objective

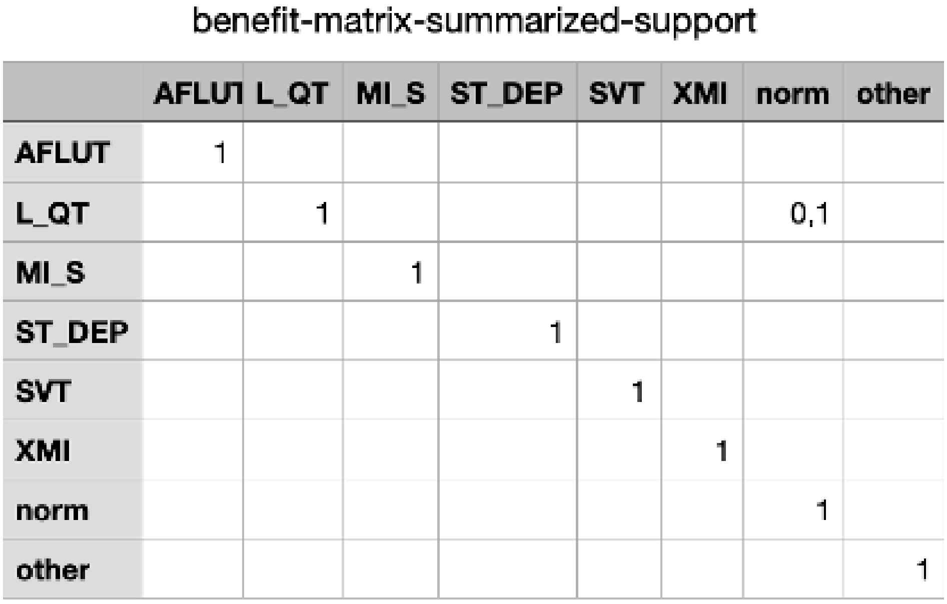
As introduced in section 5.2.1, for our *Custodian* use case, the as medically reasonable selected label compilation is characterized by different per label training/-

tuning objectives.[25] To interpret the output of a multi-label classifier in alignment with the system's intended use, and to derive domain-embedded tuning objectives for training and optimization, we deduct an exemplified interpretation of the four values that comprise the confusion matrix for the 8 labels of our experiments, as depicted in Figure 7.6. For all labels, the goal is to maximize true positives (TP) and true negatives (TN). The dataset composition reveals that nearly half the samples are healthy (norm), with approximately one-third representing unspecific pathologies, see Figure 7.1. While *norm* and *other* labels display a relatively balanced distribution of positives and negatives, other labels demonstrate a significant prevalence of negative samples (i.e. they are not present). False positives (FP) should be minimized, recognizing that some misclassifications are more tolerable than others, depending on the ground truth's medical interpretation. The implications of misclassification are particularly clear when a healthy sample (*norm*) is incorrectly predicted. False negatives (FN) are critical to avoid, as they result in missed diseases or inappropriate patient treatment. The severity of these misclassifications varies across labels and can be systematically monitored by implementing label-specific classification thresholds.

7.2.5. Development – Optimization

Optimization is closely tied to *Evaluation*. Measuring the success of experimental methods, while aiming to enhance the intelligent system is based on the chosen performance metrics, among other evaluated qualities such as fairness or generalizability. By employing metrics that accurately reflect the desired model behavior, researchers can develop a more reliable performance evaluation approach, including considerations on which metric to monitor during model training. Based on the previously defined compilation of metrics and all documented design decisions, the final step in the first cycle of metrics design is optimization. To achieve this, multiple experiments are conducted, and their structure is documented to optimize hyperparameters while considering use case-specific challenges. Focusing on ECG multi-label classification performance metrics [EM24] [EM25c], we append a domain-embedded post-processing threshold optimization method to the leaf-QG compilation, illustrating our proposed MQG4DK contribution based on CIST [LY21]. This optimization approach aims to adjust per-label thresholds that align raw output with the previously introduced four values of the confusion matrix, the foundation to calculate multiple metrics. In addition, we analyze ROC and PR calculated thresholds, complementing the fixed value 0.5. Generally, with respect to a dynamic data set with an evolving and diverse per-label distribution, threshold optimization needs to be addressed in a continuous manner along the AI lifecycle. Finally, we explore which metric to monitor during training.

[25] https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/2_Lifecycle/2_Development/2_Model_Evaluation/PerformanceMetrics/QG_Objective_(MultiLabelClassification).md



benefit-matrix-summarized-support

	AFLUT	L_QT	MI_S	ST_DEP	SVT	XMI	norm	other
AFLUT	1							
L_QT		1					0,1	
MI_S			1					
ST_DEP				1				
SVT					1			
XMI						1		
norm							1	
other								1

Figure 7.7.: Illustration of the proposed benefit matrix (BM) to consider label correlations for multi-label ECG classification in EM.

7.2.5.1. Label Correlations

As previously outlined, label correlations provide valuable insights into the intelligent system's performance within the medical domain. Understanding these relationships helps assess the model's ability to differentiate between clinically relevant conditions and informs optimization strategies. One possible approach to leveraging this information is the *Benefit Matrix* (BM) [LY21, 7], as also proposed by Physionet [26]. This method evaluates how the presence or absence of one label influences the prediction of others, thereby offering a structured way to analyze interdependencies between HARHDs. By incorporating such domain-specific correlation analysis, the system's interpretability and reliability can be improved, ultimately supporting more informed clinical decision-making.

Benefit-Matrix Label correlations are represented through a use case-specific BM[27], which captures similarities in follow-up treatments by comparing model predictions to the ground truth across all labels. In collaboration with a medical domain expert, we developed a customized BM tailored to our label structure, specifically designed for emergency medical care. The matrix values increase as follow-up steps for different labels become more aligned, with the goal of minimizing potential negative impacts on patients. These values range from 0.1 to

[26] https://github.com/physionetchallenges/evaluation-2021

[27] https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/2_Lifecycle/2_Development/3_Model_Optimization/PostProcessing/QG_BenefitMatrix_(MultiLabelClassification).md

1, where the highest value signifies complete similarity in follow-up treatment, as illustrated in Figure 7.7. To integrate this information with the intelligent system's design (as well as information on label support in the data), we apply post-processing thresholding [LY21].

7.2.5.2. Post-Processing Thresholding

Data imbalance can be mitigated through post-processing thresholding,[28] where thresholds are individually adjusted per label. The threshold applied to the model's predictions after training can be fine-tuned during evaluation. This approach is particularly useful for addressing class imbalance, ensuring that performance remains aligned with the dynamics of the relevant data distribution. If necessary, thresholds can be adapted dynamically. Alongside a fixed threshold of 0.5 for all labels, we evaluated three label-specific thresholding methods: ROC AUC-based, PR AUC-based, and CIST (with alpha set to 0.3). Each thresholding method exhibits notable differences in performance metrics, particularly for Recall and Precision. These variations become more pronounced as data scarcity increases, with the methods showing distinct augmentation behaviors depending on label support. Among these, CIST demonstrated the most balanced performance across all metrics.

Category Imbalance and Cost-Sensitive Thresholding To integrate medical knowledge on label correlations alongside support information, CIST [LY21] presents a suitable approach. The connection to domain expertise is established through the previously introduced BM, ensuring a more informed and context-aware thresholding strategy. Information on label correlations, as documented within the BM, is converted into misclassification costs, which are subsequently integrated with label support data and used to derive label-specific thresholds [LY21, 6]. The interaction between these factors is controlled by the hyperparameter *modulating alpha*, which ranges from 0 to 1. Here, class imbalance influences the thresholds with a weight of $alpha$ (0.3), while misclassification costs contribute with a weight of $1 - alpha$ (0.7), as described in [LY21, 9].

Post-Selection Analysis Evaluating CIST performance in comparison to ROC and PR curve-based thresholding, as well as fixed thresholds of 0.5 for all labels is an essential part of the post-selection process. The goal of this analysis is to identify the thresholding method that best balances both axes (Precision and Recall), represented by different metrics, while accounting for the existing data distribution. Summarizing our experiments, ROC-based thresholds exhibit the widest range of values, nearing zero for rare classes and reaching up to 0.66 for

[28] `https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/2_Lifecycle/2_Development/3_Model_Optimization/PostProcessing/QG_Thresholding_(ClassificationPerformanceMetrics).md`

norm. This method performs well primarily for prevalent labels and is suitable when high precision is the primary objective. In contrast, PR and CIST thresholds appear to better accommodate class imbalance, though the high variance in PR thresholds warrants further investigation. CIST follows a pattern where more frequently occurring labels receive higher thresholds, resulting in generally lower thresholds that enhance Recall. PR thresholds, on the other hand, tend to be higher than CIST thresholds, favoring Precision, FPR, and Specificity, which is similar to a fixed threshold of 0.5. From a macro-averaged perspective, CIST delivers the most balanced and satisfactory performance across all metrics. It achieves the highest scores for F1-Score, Recall, MCC, and Jaccard-Score. However, it is surpassed by the fixed (0.5) and PR-based thresholds in Precision, FPR, and Specificity due to its lower thresholds. A more detailed analysis is available on Zenodo[29].

7.2.5.3. Metric-to-Monitor

Additional considerations for a training/optimization strategy comprise a reasonable metric-to-monitor during model training, which is calculated on the validation data set. Choosing the appropriate metric[30] ensures that the model's performance aligns with the intended medical objectives and avoids misleading optimization results. Extending metrics selection, for further optimal hyperparameter selection[31], it is essential to optimize using a consistent error measure to enable meaningful comparisons [TJ21, 13]. This measure should be contextually embedded and well understood within the specific medical application.

Metric to Monitor Adjusted to our fictional use case, we compare F1- and FBeta-Scores within our experiments, monitoring early stopping and learning rate adjustment. These considerations are based on the previously introduced use case-adapted training/tuning objective, which emphasizes the interplay between Precision and Recall. Following our domain-embedded approach, the selected metric-to-monitor best represents the real world objective we aim to achieve. Our experiments demonstrate that this design decision influences performance metrics in a highly granular way. With a fixed threshold of 0.5, micro- and macro-averaged results were nearly identical, as the model's confidence tended to be relatively low. However, with lower CIST ($alpha$ = 0.3) thresholds, FBeta-Score favors Recall for both averaging methods, while F1-Score shows a more balanced performance improvement of a few percentage points across the other confusion

[29] `https://zenodo.org/records/10931084`

[30] `https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/2_Lifecycle/2_Development/2_Model_Evaluation/PerformanceMetrics/QG_PerformanceMetricsCompilation_(MultiLabelClassification).md`

[31] This is not the focus of our analysis and comprises design decisions such as the loss function, learning rate (reduction), or the inclusion of early stopping.

matrix-based metrics.[32] ROC AUC and PR AUC performances were almost identical for both metrics-to-monitor. In general, the performance values for both metrics-to-monitor were similar, but macro-averaging displayed a higher variance (greater than 10%) compared to the micro-averaged metrics (approximately 10%).

Impact on Thresholding With respect to the calculated thresholds, the following behavior was observed as a consequence of the metric-to-monitor:

- CIST: When the F1-Score is monitored, a slightly higher threshold is set for rare classes than when using FBeta. For majority classes, the threshold is marginally lower, which aligns with the fact that F1-Score prioritizes Precision more.

- PR: There is significant variance in the direction and magnitude of threshold changes caused by the metric-to-monitor, likely because of the balanced emphasis on Recall and Precision.

- ROC: When FBeta-Score is monitored, the threshold is typically higher, possibly due to the focus on Recall. However, the overall difference is very small.

Concluding our analysis, a reliable metric selection forms the foundation for further tuning, as well as for monitoring (and deployment) of the intelligent system within its intended real world environment, and it is closely linked with other design decisions. This process also involves planning for when a re-evaluation is necessary. In our experiments, we focused on two key design decisions for optimization, that are embedded within our ECG use case and impact performance metrics. We aim to extract generalizable guidelines for the metric-to-monitor and post-processing threshold optimization for the multi-label case, while keeping other hyperparameters unchanged.

After introducing relevant considerations along the AI lifecycle, focusing on development, the next section outline MQG4AI-related design guidelines, addressing resulting QG types, including relevant information. As previously introduced, multi-label and binary classification metrics are calculated in a similar manner, resulting in *QG Inheritance* within our contribution to MQG4DK. In addition, we briefly highlight the *QG Naming* and *QG Tags* structure for the coherent QG-compilation.

[32] This outcome aligns with their respective calculations, with the FBeta-Score valuing Recall twice as important as Precision, compared to the F1-Score.

7.2.6. *Leaf-QG* Relations

In chapter 6, we exemplified the leaf-QG information processing template for a technical guideline, as introduced in section 4.3. Analogously, starting from development, relevant information for the subsequent maintenance and deployment stages is extracted based on the leaf-QG information layers. Thus, linking lifecycle stages, risks and system knowledge is included by design, as can be viewed in more detail on GitHub[33]. In this section, with respect to the proposed compilation of leaf-QGs, we highlight emerging leaf-QG relations, as additional important information on how to interact with MQG4AI. Concretely, we shed light on relations emerging from similar content, as with binary and multi-label classification, and briefly highlight the exemplified QG naming and QG tags that enable IM within MQG4AI.

7.2.6.1. QG Inheritance

The *QG Inheritance* relationship emerges due to the structural similarity between binary and multi-label classification. Basically, the latter is calculated through the addition of averaging methods. With respect to per-class probabilities, the binary calculation takes place, only with a slightly different meaning (i.e. the presence or absence of a label). As introduced in section 4.1.1, we establish the *QG Inheritance* relationship when similar information is extended, derived from object-oriented programming (OOP). This setup may be transferable to other design decisions for organization within MQG4DK. Further, more possible *QG Relations*, e.g. based on OOP software engineering principles such as *aggregation* or *composition* [Par20] may emerge between QGs that append more design knowledge to MQG4DK, which is part of future work.

7.2.6.2. QG Naming

QG Naming within the context of MQG4DK indicates the content of the respective leaf-QG contribution in the format *QG_name_(view)*. View refers to the applicability of the individual contribution. Within our proposed QG compilation for reliable multi-label ECG classification performance evaluation metrics in EM, *multi-label classification* or *classification performance metrics* comprise the most prevalent views, aiming to highlight related design decisions in form of leaf-QGs. Regarding the *QG Inheritance* relationship, the two leaf-QGs share the same *name*, while the *views* differ:

[33] `https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/3_RiskManagement/AI_Risks/2_TechnicalRobustnessSafety/Accuracy/UnreliablePerformanceMetrics.md`

- *QG_PerformanceMetricsCompilation_(Classification)*

- *QG_PerformanceMetricsCompilation_(Multi-LabelClassification)*

7.2.6.3. QG Tags

QG Tags organize information when pulling MQG4A-v0 from MQG4DK. They aim to support the extraction of strictly relevant QGs for a concrete application scenario through an envisioned intelligent search and are based on ML design pattern [WH22]. The following information organization tendencies comprise our exemplified contribution, highlighting the connected leaf-QG compilation character:

- *Name* Refers to *name* of the previously introduced *QG naming* structure, providing information on the respective design decision.

- *Intent* The intent is summarized as *EvaluationStrategy*, *PerformanceOptimization* or *AdditionalMaterial*, for instance, mirroring the individual QG's functionality.

- *Problem* The problem section references risks such as *Data Imbalance* or *Unreliable Performance Metrics*.

- *Solution* provides information on a desirable outcome, which is enhanced by the individual QG. This comprises technical information such as *Error Quantification* when transforming raw model output based on the confusion matrix, *Use Case-adapted Label Correlations* when applying the benefit matrix, or the *Iterative Design Decision-making Method* when choosing the performance metrics compilation.

- *Applicability* refers to structural settings where the individual QG is applicable. In our case, this includes *(Multi-label) Classification* and may append further information such as *Thresholding* for the benefit matrix, for instance.

- *Consequences* summarize the result when applying the leaf-QG. For instance, implementing a groundtruth-probability distribution along with the performance metrics compilation leads to a *Visualization*, which equally supports metrics *Monitoring*. Additionally, post-processing thresholding results in adjustable thresholds, equally related to metrics *Monitoring*.

- *Usage Example* highlights what all QGs share, the fictional *Custodian* ECG use case situated in EM.

7.3. Evaluation

This chapter illustrates a complex contribution to MQG4DK in the form of a leaf-QG compilation towards reliable performance evaluation metrics for multi-label ECG classification, aligned with a fictional use case situated in EM, highlighting medical domain knowledge. In this section, we reflect on the proposed approach. We focus on the evaluation lifecycle stage during model development and contribute a combination of relevant related design decisions in form of QGs. Concretely, the leaf-QG representation layer was not focus of our experiments, as well as the propagation of how to integrate the performance metrics compilation with other subsequent lifecycle stages, such as deployment and maintenance.

Evaluation Lifecycle Design The interdisciplinary and domain-adapted approach, as previously introduced, along with the derived guidelines for multi-label (ECG) performance metrics, contribute to the development of a TAI lifecycle. Instead of focusing on model optimization, we intend to identify empirical trends. Given that reliable metrics are a key aspect of general AI RM [Tab23, 6], we emphasize the importance of establishing a solid foundation from the outset. An additional promising direction, though beyond the scope of our study, is the development of a custom evaluation metric in close collaboration with domain experts. This approach could refine existing metrics to better align with specific application needs. For an example, refer to the Physionet Scoring Metric[34], as further possible related contributions to MQG4DK. This design decision may be appended to MQG4DK as an additional leaf-QG, for instance, complementing the proposed setup, and linked through QG naming and QG tags. Overall, our experimental approach to multi-label ECG classification is not the optimal solution. Depending on the specific tuning objectives of each label, alternative methods, such as ensemble learning, could be more effective. These aspects should be addressed during a thorough pre-selection phase, where multiple methods are initially evaluated in parallel to determine the most suitable approach. Additional refinements include a different approach to metrics selection. Highlighting stakeholder inclusion and further design techniques, open questions centered around use case-specific lifecycle design for model evaluation include:

- A more profound integration of clinically distinct tuning objectives per class. For instance, by adjusting the balance between Precision and Recall, the trade-off between false alarms and missed diagnoses can be tailored more specifically to each label.

- A revised BM for comparison and a potentially more realistic adaptation of domain knowledge. This could involve reassessing the initial approach, incorporating clinical prioritization of class labels for improved alignment with real world applications.

[34] https://moody-challenge.physionet.org/2021/

- Relationships with other (existing) Quality Guidelines in form of QGs must be identified and aligned more comprehensively and possibly with further experimental testing.

- XAI techniques may enhance the interpretation of model outputs, metric performance, and supplementary materials.

- Additional metrics should be considered for a comprehensive evaluation strategy, such as those assessing fairness, as outlined in ISO 24027 [ISO21a].

- Adapting relevant results for stakeholders throughout the AI lifecycle is essential for post-market surveillance and usability. This may include onboarding strategies to help (medical) users interpret metrics, through e.g. interactive monitoring dashboards.

- For auditor representation, a structured textual template could be developed based on the proposed approach, incorporating use case-specific details and referencing supplementary material with in-depth results for concise conformity assessment.

- A dedicated metric could be designed to monitor the performance and reliability of other metrics for enhanced surveillance.

- Benchmarking would benefit from established metrics compilations for (groups of) use cases. Also, comparing performance against the current clinical state-of-the-art (including already applied evaluation metrics) is essential for optimization and for validating the integration of AI in medical applications.

Guiding Questions: Aiming to guide developers within concrete MQG4A scenarios to transfer the design knowledge pulled from MQG4DK to their respective use cases, *Guiding Questions* serve as a suitable method. We propose to present guidelines in form of questions to extract information following the *Socratic Method* [IA12], as introduced in subsection 4.3.1. The insights derived from our experiments can be structured as follows to emphasize the integration of trustworthy metrics into the AI lifecycle from the outset. These guidelines are designed to support reliable long-term monitoring and contribute to RM. With a focus on both, developers and potentially auditors, examples to begin lifecycle design early on include:

- What clinical impact is intended or caused by the model's output?

- Which metrics are relevant to different stakeholders at various stages of the AI lifecycle?

- What additional stakeholder perspectives need to be considered (e.g. patients)?

- Who has a global overview and is responsible for ensuring proper guidance? What other roles and representatives are involved in accountability management?

- How should metrics (including supplementary information) be visualized based on their intended application?

- At which stages of the intelligent system's development and depending on which parameters, should decisions be monitored and potentially adjusted?

- What interdependencies exist between other design decisions or lifecycle stages (represented by QGs)?

MQG4DK Contributions Aiming to enable more MQG4DK contributions, the question how to organize them emerges, especially, with respect to QG naming and QG tags. Highlighting concrete use cases, an ontology of AI application scenarios within a particular domain such as medicine may be reasonable, in addition to considering structural similarities. As a result, QG naming and QG tags need to be analyzed and tested further to enable complete AI and use case coverage. Therefore, use cases need to be collected, in alignment with domain knowledge, an then summarized according to a combination of criteria. For instance, relevant characteristics for use case categorization may address the structural setting such as the model's functionality, and domain-embedded tuning/training objectives. As introduced in section 5.1, for instance, ISO/TR 24291:2021(E) on *Applications of machine learning technologies in imaging and other medical applications* [ISO21b] provides a framework for identifying, classifying, and evaluating intelligent medical applications. Additional considerations comprise the high-level setup of the proposed lifecycle development stage. Within our leaf-QG contribution, even though focusing on performance metrics evaluation, other stages, such as optimization, are equally relevant. More concrete examples and contributions to MQG4DK are required to assess the lifecycle development stage proposition for broad applicability and to derive a generalizable and trustworthy collection-QG structure.

After focusing on MQG4DK contributions in the previous chapters, the next chapter introduces a simulated MQG4A scenario, exemplified for segmentation model selection embedded within our interdisciplinary project on the rare disease Achalasia, as introduced in section 5.2.2.

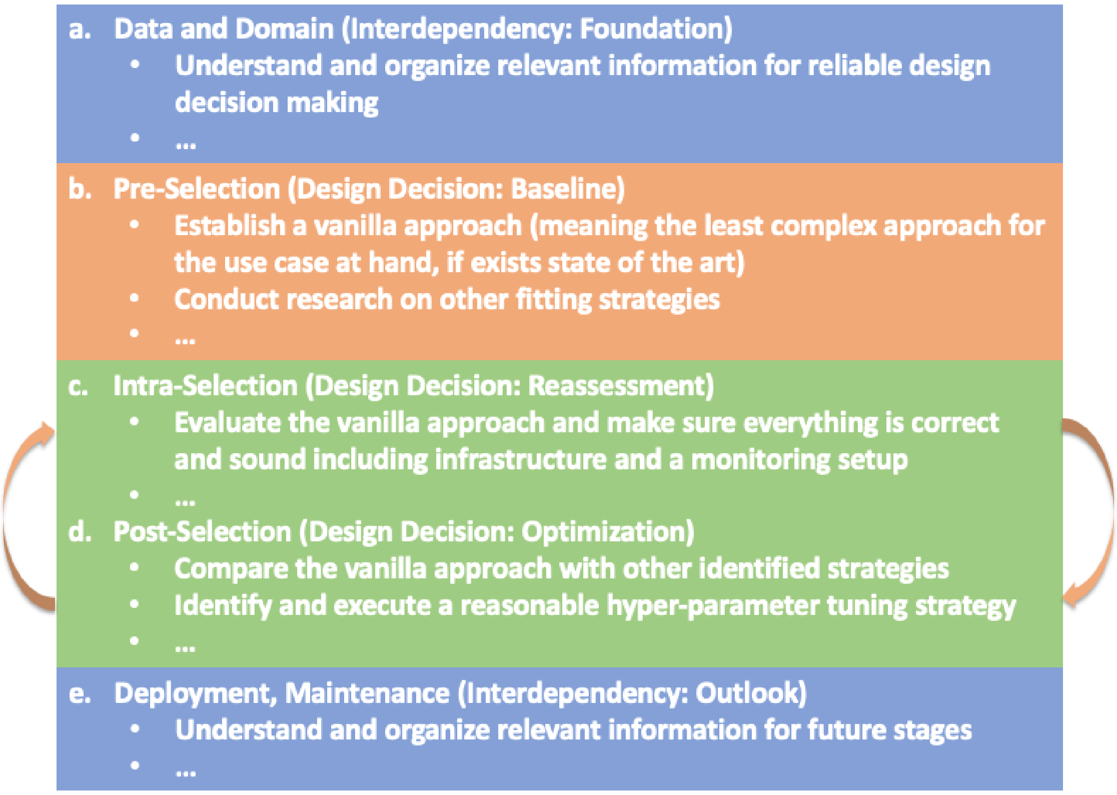
8

MQG4A Template Versions – QG *Model Configuration* for Timed Barium Esophagogram Image Segmentation with a Human-in-the-Loop

This chapter simulates a model selection scenario during the development stage based on the supervised bachelor's thesis *Oesophageal Segmentation using Barlow Twins and Multi-Label Annotation in the context of the Rare Disease Achalasia - A Self-Supervised Learning Approach*, as introduced in section 1.6.3. The corresponding use case, centered around the tool *EsophagusVisualization*, is introduced in section 5.2.2. In short, the segmented model output is embedded within the medical software, and enhances a workflow step to design a 3D-reconstruction of the esophagus, which includes the medical human-in-the-loop. This section is to be interpreted in combination with the illustrated MQG4AI lifecycle blueprint on GitHub[1]. Our overall objective is to present a clear and accessible MQG4A scenario, with the emphasis placed on conceptual understanding rather than the underlying technical implementation. On an abstract level, we compare two different approaches to choosing the model configuration and link relevant information with QGs along the AI lifecycle, embedded within additional information on the AI system, including identified risks.

Through this retrospective analysis, we aim to introduce MQG4A information organization in form of different template versions following pre-, intra-, and post-selection information clustering for reliable design decision-making [EM24] [EM25c], as illustrated in Figure 8.1, and introduced in the previous chapter 7. We focus on the lifecycle development stage, and aim to illustrate the reasoning why a particular model configuration is chosen in a simplified manner. This intends to highlight the MQG4A mechanism of instantiating and utilizing different template versions for RAI IM and KM. As introduced in section 4.2.1, monitoring the

[1] https://github.com/miriamelia/MQG4AI/blob/main/README.md



a. Data and Domain (Interdependency: Foundation)
- Understand and organize relevant information for reliable design decision making
- ...

b. Pre-Selection (Design Decision: Baseline)
- Establish a vanilla approach (meaning the least complex approach for the use case at hand, if exists state of the art)
- Conduct research on other fitting strategies
- ...

c. Intra-Selection (Design Decision: Reassessment)
- Evaluate the vanilla approach and make sure everything is correct and sound including infrastructure and a monitoring setup
- ...

d. Post-Selection (Design Decision: Optimization)
- Compare the vanilla approach with other identified strategies
- Identify and execute a reasonable hyper-parameter tuning strategy
- ...

e. Deployment, Maintenance (Interdependency: Outlook)
- Understand and organize relevant information for future stages
- ...

Figure 8.1.: This methodological structure is derived from empirical insights on reliable performance evaluation metrics [EM24] [EM25c], and we transfer this three-fold structure to simulate model configuration selection in form of MQG4A template versions. As will become clear in the following sections, *a* and *b*-related, as well as *d* and *e*-related information may overlap depending on the respective design decision and template version within the overall lifecycle design workflow.

resulting template structure is part of the *Conceptualization* lifecycle stage and envisioned as an additional management layer during project execution. Therefore, in contrast to MQG4DK contributions, MQG4A scenarios include concrete results within the content-layer of leaf-QGs, as detailed in sections 4.3 and 6.2.2. The envisioned MQG4A scenario to design and plan the project-specific AI lifecycle is depicted in Figure 8.2.

Generally, template v1 builds on v0, which is pulled from MQG4DK. First, system-related information is appended, that for instance, describes the application and relevant stakeholders. In addition, related risks are identified, based on which lifecycle design decisions (QGs) for mitigation are identified, possibly resulting in a lifecycle design roadmap. The subsequent template versions continue to append lifecycle design decisions in a tested manner, and possibly novel related risks to the *main* lifecycle implementation. This perspective focuses on the case when multiple design candidates are tested for the final implementa-

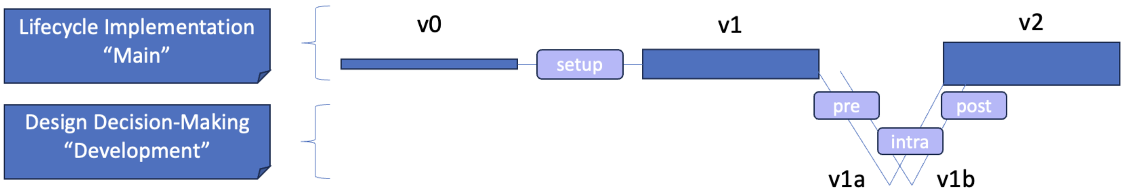


Figure 8.2.: A proposition for MQG4A versions. After pulling v0 from MQG4DK, and filling in system information, including related risks, lifecycle design-decision-making begins, resulting in multiple sub-steps that grow the *main* version. In addition to lifecycle design decisions, this may comprise novel related risks and other contextual information.

tion. Lifecycle conceptualization is realized through template snapshots that are centered around specific lifecycle design decision-making, development or design decision *testing* (referring to Git branching structure terminology)[2], aiming to document the path to building the *main* (or, *master*) lifecycle concept for validation and KM. In theory, these comprehensive and continuously growing MQG4A template versions are accessible by all contributing stakeholders, and a reasonable access distribution is identified in alignment with the individual project. For instance, during compliance assessment, auditors may require checking some design decision-making, or *testing* steps.

The following sections further describe the generated template versions within our application scenario for choosing a *model configuration* for timed barium esophagogram (TBE) segmentation within the medical software *EsophagusVisualization* based on the human-in-the-loop approach. Section 8.1 comprises preselection information, and first outlines the system-information block, as well as illustrates RM centered around the model's intended use. Next, we introduce the predefined metrics compilation and data design input within our retrospective simulation of MQG4A for *model configuration* selection. The thus generated information comprises the foundation for the, in section 8.2, outlined intra-selection template versions comparing two methods for segmentation model selection based on concrete results. In section 8.3, merging the chosen *model configuration* method with the *main* lifecycle template, and appending additional relevant QGs for post-selection, is illustrated. Finally, focusing on the development stage, in section 8.4, we evaluate this simulation of MQG4A based on constraints that emerge from its retrospective character, in addition to our use case that is centered around a first prototype for medical research, developed as part of higher education.

[2] https://git-scm.com/book/en/v2/Git-Branching-Branches-in-a-Nutshell

8.1. *Main*-Template (v1): Foundation & Pre-selection

Additional lifecycle design decisions are required previous to *model configuration* intra-selection steps, and they extend beyond AI system-related information and identified risks. Among others, additional pre-selection phases comprise the data design input during the data acquisition stage, as well as a pre-defined performance metrics compilation for model evaluation. Therefore, within our MQG4A simulation, v1 is not actually v1 of the complete lifecycle design workflow. As a result, v1 within our retrospective analysis comprises system, and RM information, as well as lifecycle QGs, as introduced in the following sections.

8.1.1. System Information

The system information block, as introduced in section 3.3.3 is filled for the Achalasia use case and *EsophagusVisualization* tool, as introduced in section 5.2.2.

1. *Application* provides information on the medical software, and how the segmentation model is embedded within the 3D reconstruction workflow. This comprises the construction algorithm, highlighting the role of the data input (TBE), as well as the overall research project vision. In addition to contextualizing the TBE segmentation model, *Domain Knowledge* on the rare disease *Achalasia* is provided within the lifecycle template.

2. *Documentation* In addition to documenting the lifecycle design through MQG4A template versions, we append a user manual, as well as instructions for installing the software, focusing on how to integrate the segmentation model.

3. *Stakeholder* For our simulation, we append three stakeholder roles. *Active* roles comprise the model *developer*, in addition to a *Consulting physician* during model development, highlighting the need for annotated data. Finally, a *Passive* role is identified, focusing on the *physician* who interacts with the model during operation.

4. *Ethics* Information on *Ethics_General* and *Ethics_Specific* was provided to our project team through regular discussions of e.g. the importance of domain embedding and stakeholder inclusion during development, and the role of rare diseases in society. No particular ethics training was conducted.

No additional resources were considered for model selection, and the other system information modules are not altered within our simulation, and reflect MQG4DK.

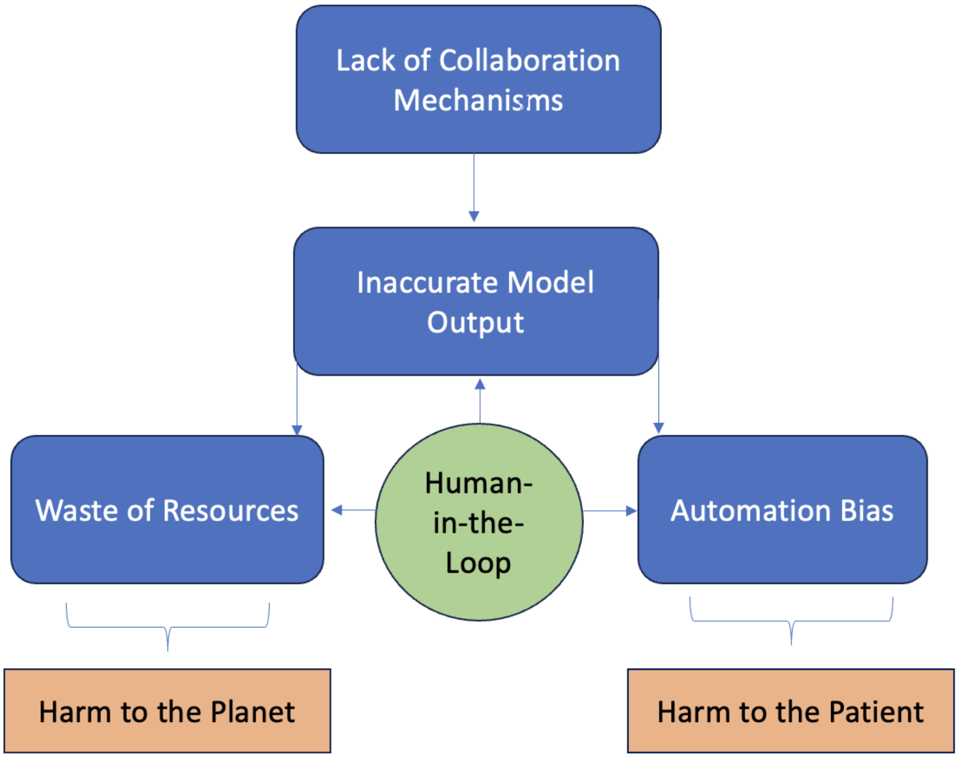


Figure 8.3.: Illustration how a combination of risks are organized, related to *inaccurate model output* and consequences focusing on the medical user. An unfitting *model configuration* is a risk source for *inaccurate model output*. Additional risks such as data quality-related risks are not depicted.

8.1.2. Risk Management Information

In our simulation, we focus on the physician's interaction with the intelligent system, considering risks associated with model output quality that are linked directly to the model configuration. If not mitigated, these risks could compromise the system's *intended use*, i.e. assisting physicians in generating 3D reconstructions of the esophagus through an optimized software workflow. The following identified risks are organized in accordance with the TAI structure, as introduced by ALTAI [Hig20]. The key risk *inaccurate model output* is directly linked to the model configuration, based on which we derive three additional risks (one source and two consequences) that impact RAI lifecycle design, centered around the human-in-the-loop approach for model application within the medical software. Additionally, other risks beyond model output quality include data quality in general, privacy concerns, and robustness-related security threats, such as data poisoning or model evasion. Further risks may emerge as design decisions evolve, documented through evolving template versions, which all contribute to the *main* lifecycle planning version.

8.1.2.1. Inaccurate Model Output

The risk *inaccurate esophagus segmentation* may cause medical users to forgo using the segmentation model within the software or require extensive manual adjustments, leading to a more time-consuming 3D reconstruction workflow (related risk: waste of resources), or an inaccurate foundation for the 3D reconstruction of the esophagus (related risk: automation bias). It is documented under *Technical Robustness and Safety*, *Accuracy* within the MQG4A template structure.

1. *Risk Sources* Insufficient segmentation accuracy may arise from various factors, including limited labeled data during model development, poor data quality, an unsuitable model configuration, or data and model drift over time during maintenance. We derive two additional risks focusing on how physician's interact with the lifecycle based on these sources. Namely, a lack of collaboration mechanisms and, indirectly, domain expert input evaluation with respect to data labeling quality.

2. *Risk Analysis* The *severity of harm* associated with this issue is that the 3D reconstruction workflow may not be optimized, leading to inefficiencies. As a result, physicians interacting with the software may choose not to utilize the TBE segmentation model, ultimately causing a waste of resources, or, incorporating *automation bias*, they may accept the inaccurate output, leading to inaccurate 3D reconstructions, and a negative impact on patient well-being. The *probability of occurrence* is high, particularly in the early stages of the project, when the model and its integration are still being refined.

3. *Risk evaluation* The risk is considered acceptable if appropriate control measures are implemented and updated. Without these measures, the model may hinder software application, or not be applied at all in practice, rendering its implementation ineffective, and a waste of resources.

4. *Risk control* To mitigate this risk, several control measures should be implemented. In cooperation with the consulting physician, an acceptable segmentation performance threshold needs to be defined that ensures model application. In addition, collecting more annotated data and retraining the model will enhance segmentation accuracy. Additionally, optimizing the model configuration can further refine its reliability and effectiveness. To ensure continuous monitoring, a feedback mechanism should be established, allowing physicians to report any observed declines in model performance. This will facilitate ongoing improvements and help maintain the model's applicability. Other approaches include high-performing hardware that enables GPU usage during inference, as well as the option to deselect segmentation model application from the treating physician's view for e.g. out-of-distribution samples with low model performance.

8.1.2.2. Lack of Collaboration Mechanisms

A *lack of collaboration mechanisms* may cause inaccurate model output, since physicians interacting with the model during operation (maintenance lifecycle stage) are unable to report performance declines, potentially leading to the model not being applied, which results in a waste of resources. Within the template, this risk is situated under *Stakeholder Participation*, which belongs to *Diversity, Non-discrimination, and Fairness*.

1. *Risk Sources* Resulting from a lack or malfunctioning of feedback channels for physicians, collaboration mechanisms are not established. Further, if the physician in charge does not know how to use them, positive effects of collaboration mechanisms are unsuccessful.

2. *Risk Analysis* Without collaboration mechanisms, physicians may be unable to report performance declines, resulting in inaccurate model output. Thus, *severity of harm* and *probability of occurrence* are as described above.

3. *Risk evaluation* This risk is deemed acceptable if appropriate controls are in place that ensure a functioning communication infrastructure.

4. *Risk control* To address the risk of lacking collaboration mechanisms, it is essential to establish effective feedback channels during system operation that enable communication between the physicians interacting with the software and the development team. Within the context of *EsophagusVisualization*, which is developed in an interdisciplinary manner in the context of higher education and research at the University of Augsburg and University Hospital Augsburg, collaboration mechanisms are established for the first prototype in form of in-person and digital meetings. However, once the application is distributed among multiple centers, implementing structured feedback loops will globally allow physicians to communicate observations, ensuring continuous model improvement and sustained clinical applicability.

8.1.2.3. Waste of Computing Resources

A *waste of computing resources* comprises a direct follow-up risk of inaccurate model output, if, depending on the physician, it may be preferred to not apply the model and insert the esophagus' shape manually instead. Thus, its development and training efforts are ineffective, which leads to wasted computing resources. Within the template, this risk is organized under *Environmental Well-being*, which comprises a subsection of *Societal and Environmental Well-being*.

1. *Risk Sources* These are analogously to inaccurate model output, since this is a derived risk, which impacts the estimation of the severity of occurrence regarding inaccurate model output.

2. *Risk Analysis* The *severity of harm* is measured regarding the utilized computing power in light of an inefficient use of computational resources due to an underutilized model, which harms the planet. The *probability of occurrence* is highly dependent on both, the model's performance and the physicians' assessment of its reliability. It is more likely in early deployment stages or if no mechanisms are in place for continuous improvement, which includes the establishment of collaboration mechanisms.

3. *Risk evaluation* The risk acceptability is contingent on physician approval, which requires regular validation to ensure continued use.

4. *Risk control* Options for risk mitigation align with the risk of inaccurate model output, as previously introduced. They comprise ongoing model updates and refinements, as well as early detection of performance degradation that triggers necessary adjustments, e.g. through feedback loops.

8.1.2.4. Automation Bias

The risk *automation bias* comprises a possible follow-up risk that results from TBE segmentation model misuse. If the physician in charge who interacts with the software relies too heavily on the generated mask, without verification or corrective adjustments to the segmented esophagus' shape. It is assigned to *Human Oversight* within *Human Agency & Oversight*.

1. *Risk Sources* Inaccurate model output comprises a risk source, in addition to the human-in-the-loop.

2. *Risk Analysis* The *severity of harm* is evaluated based on the effects of the inaccurate esophagus' shape that reduces the accuracy of the generated 3D reconstruction, which could negatively impact medical research and, in the future, patient well-being. The *probability of occurrence* depends on the physician's level of engagement and awareness when using the software.

3. *Risk evaluation* Since physicians are trained to recognize the esophagus in TBE images, they are likely to adjust inaccuracies as needed. Additional risk controls further support acceptability of this risk.

4. *Risk control* Clearly document the limitations of the segmentation model to ensure physicians remain aware that manual corrections may be necessary.

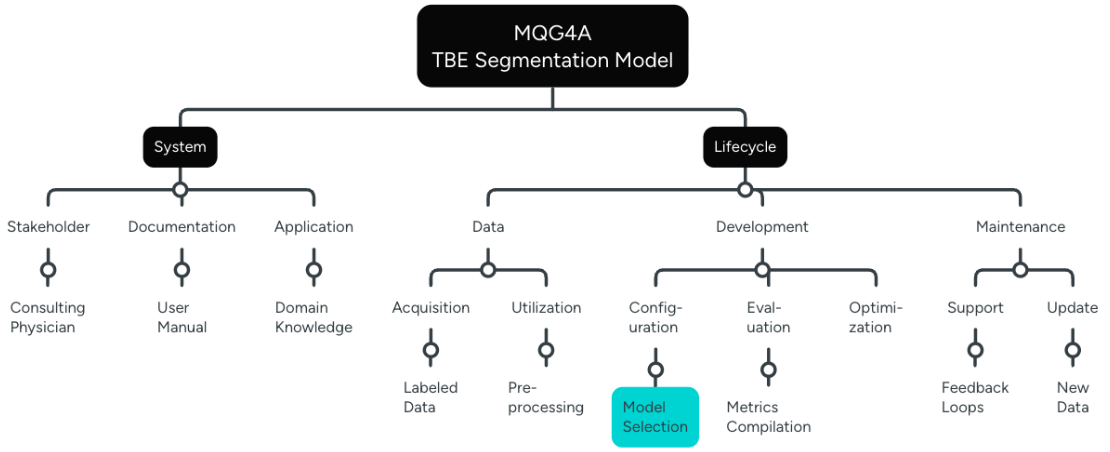


Figure 8.4.: QGs and template information related to mitigating the risk *inaccurate model output*. This lifecycle snapshot is depicted without information on the deployment phase, RM and ethics.

In addition, an equally introduced risk control for *inaccurate model output*, the software provides an option for physicians to skip the model's segmentation output if they determine it is unreliable and start a new segmentation.

The overall resulting leaf-QG compilation, as well as system information to mitigate the key risk *inaccurate model output* is depicted in Figure 8.4. The next section contributes model configuration-related lifecycle leaf-QGs to the pre-selection template. Namely, we outline data-related considerations and briefly explore suitable segmentation performance evaluation metrics.

8.1.3. Lifecycle: Metrics Compilation & Data Design Input for TBE Segmentation

In addition to RM and system related information that comprise the foundation for overall lifecycle design, related lifecycle design decisions in form of QGs needs to be identified. This should happen prior to choosing a model configuration and be embedded within the generic lifecycle structure, as introduced in section 4.2. Within our simulation snapshot, this pre-selection comprises information on input (and output) data design (TBE images, and additional segmentation masks), as well as a performance evaluation metrics compilation that enables choosing the best segmentation model selection. The two resulting leaf-QGs are introduced in the following. In addition, this stage comprises research on possible model configuration candidates that are implemented and tested as part of intra-selection.

8.1.3.1. QG Data Design Input

The leaf-QG *Data Design Input* belongs to the *Data Acquisition* lifecycle stage. It is relevant for model configuration design, since it shapes the input and output of the segmentation model. The leaf-QG sections are simulated as follows:

- *Input Information* comprises information on **Data Storage** as an additional lifecycle QG that equally belongs to the generic stage data acquisition. In addition, system-specific input information refers to a **Consulting Physician**, defining a stakeholder requirement for data labeling.

- *Output Information* Data design defines the **Data Collection**, which equally belongs to data acquisition. In addition, the design input comprises the foundation for **Data Preprocessing**, which is assigned to the collection QG data utilization within the template. Further, with respect to post-market monitoring, the data design shapes new data and possible model updates.

- *Content* For the use case at hand, the data comprises medical raw data in form of TBE images as model input, in addition to corresponding annotated segmentation masks to build a development data set. This equally corresponds to the desired model output, which needs to be integrated with the *EsophagusVisualization* tool.

- *Method* For the TBE segmentation scenario, data design is implemented in cooperation with the consulting physician. Additional technical considerations related to the implementation may also be outlined at this point, but were omitted as the tool is only a prototype.

- *Representation* From the viewpoint of our snapshot, focusing on model development, data design is relevant to **Developers** (and shaped in alignment with physicians who provide medical domain knowledge).

- *Evaluation* Finally, with respect to data design, open questions may address a multi-label annotation setup, for instance, which could enhance model performance with an augmenting amount of annotated data. Therefore, *EsophagusVisualization* allows for exporting multiple annotations for future optimization. They comprise the esophagus, the spine, as well as the barium, which may overlap, resulting in one pixel possibly belonging to multiple labels.

Finally, for instance, related risks are centered around the quality of the annotated (esophagus') masks, relating to topics surrounding inter-annotator agreements [BN20] as risk mitigating measures. Overall, we exclude related RM, since it is beyond the scope of our snapshot.

8.1.3.2. QG Performance Metrics Compilation

The second related pre-selection lifecycle leaf-QG refers to a performance evaluation metrics compilation that is suitable for the esophagus segmentation use case. This leaf-QG belongs to the *Development Model Evaluation* lifecycle stage.

- *Input Information* generally refers to the **Data Utilization** stage and comprises e.g. information on data distributions, which impact the applicability of performance evaluation metrics. In addition, **Developer** with knowledge on performance metrics and how to implement them for the use case at hand are required.

- *Output Information* within the scope of our snapshot comprises further substages of lifecycle development, highlighting the **Model Configuration** and **Model Optimization**. Regarding **Post-market-monitoring**, and the maintenance stage, the metrics are relevant as soon as the model is retrained, and model updates are realized. As previously introduced, retraining the model may be triggered based on feedback-loops by the physician interacting with the software, and/or more annotated data.

- *Content* Four metrics are selected, aiming to capture the model's performance, highlighting the relation between the predicted segmentation and annotated mask. They all range from 0-1 and should be maximized. The **Dice-Score (or F1-Score)** comprises the primary metric, which is used in most publications for medical image segmentation and describes the harmonic mean of sensitivity and precision [MD22, 6]. Precision evaluates how many of the, as esophagus segmented pixels actually belong the annotated mask. **Sensitivity (or Recall/True-Positive Rate)** evaluates the model's capability to detect the region of interest, namely, it answers the question how many of all pixels that belong to the esophagus mask the model identified. It is "[...] less sensitive to F-score based metrics for exact evaluation and comparison of methods" [MD22, 6]. Further, the **Intersection over Union (IoU) (or Jaccard Score, equally described as an F-Score metric** [MD22, 6]) is selected, which higher "[...] penalizes under- and over-segmentation [...]" [MD22, 6], describing cases when the model either missed a large amount of the target annotation, or included to many background pixels within its predicted mask, respectively. In addition to **Sensitivity (or Recall/True-Positive Rate)**, **Specificity** is calculated, as two widespread metrics in the medical domain. Specificity evaluates how well background pixels are predicted, and therefore can be interpreted as rather assessing the model's functionality in contrast to its performance. It is usually expected to be close to 1. [MD22, 6]

- *Method* The metrics compilation is selected based on a publication focusing on evaluation metrics, particularly tailored to medical image segmentation [MD22]. An important selection criterion comprises metrics that are suitable for high data imbalance, since pixels that identify the esophagus are comparably less prevalent than background pixels. Therefore, true positives and true negatives should not be weighted equally [MD22, 2].

- *Representation* A comprehension of how to interpret the selected performance metrics compilation is relevant to **Developers** from the perspective of our snapshot during model development. The physicians who interacts with the segmentation model does not need to consult metrics that evaluate overall model performance, since they directly consider the segmented masks, reducing interaction complexity.

- *Evaluation* Visualizing the segmentation model output contributes additional information on the model's performance. Possibly, more metrics could be appended to the illustrated metrics compilation within our simulation, based on [MD22].

For instance, related risks are centered around the suitability of chosen metrics. Among other risk mitigating measures, unreliable performance evaluation metrics are mitigated by information provided in the leaf-QG method dimension, where it states how they were selected. In addition, they need to be interpreted in an adequate manner, which, in our scenario highlights developers as main target group. Overall, we exclude related RM, since it is beyond the scope of our simulation. Further, the model output on the test set equally sheds light on segmentation model performance, which contributes to performance metrics interpretation and is documented within the leaf-QG evaluation layer, as illustrated in section 8.2 with concrete results. Finally, rounding up the pre-selection stage within our simulation, a brief analysis of possible TBE segmentation model configuration approaches is conducted, resulting in the two, in the following introduced methods and *development* MQG4A template versions for testing.

8.2. *Development*-Templates: Intra-Selection

Building on the previously introduced pre-selection lifecycle QGs and fundamental system considerations, the two intra-selection versions (v1a and v1b), or *development* templates for testing, each reflect a different segmentation model approach. In addition to the model configuration, they comprise design-decision-specific QGs, which include related preprocessing steps, as well as calculated performance evaluation metrics as foundation for design decision-making. These template versions highlight design decision-related QGs during model development without further information linking, as included in v1. Thus, showcasing

MQG4AI's customizability through reduced template snippets for an optimized MQG4A interaction. Comprehensive information linking with other lifecycle interdependencies is continued in v3, post-selection, in addition to appending the design decision on which segmentation model configuration to apply to the *main-branch*. Analogously, other configuration approaches can be appended. This section briefly outlines two different approaches to TBE segmentation model configuration, aiming to illustrate how to approach comparing two possible design decisions against one another within MQG4A. Namely, one approach, as found in the literature for high data imbalance, is based on the RCA-IUnet [PA22c] and Barlow Twins for self-supervised learning [PA22b]. We compare this method against the state-of-the-art nnU-Net [IF21]. The results clearly indicate that, with our limited amount of annotated TBE images, the nnU-Net outperforms the alternative approach, and it might be reasonable to rerun these experiments, as labeled data augments.

8.2.1. v1a: State-of-the-Art & Benchmark (nnU-Net)

The "no new" U-Net (nnU-Net) is a DL-based segmentation framework that autonomously configures key components such as preprocessing, network architecture, training, and post-processing for new tasks. Its design is guided by a combination of fixed parameters, interdependent rules, and empirical decisions, allowing it to adapt effectively without manual intervention, based on a "data fingerprint" [IF21, 205]. Notably, it has demonstrated superior performance over existing approaches, including highly specialized solutions across multiple public datasets from international biomedical segmentation challenges. In addition, the code is publicly available.[3] Therefore, given its ability to generalize across diverse tasks while maintaining state-of-the-art performance, nnU-Net was chosen as the benchmark segmentation approach for our simulation. [IF21]

8.2.1.1. QG Data preprocessing

Preprocessing is carried out automatically by the nnU-Net framework. A comprehensive overview of possible methods can be found in the supplementary material on nnU-Net design principles of the corresponding publication [IF21, 205].

- *Content* In total, the labeled data set consists of 85 TBE images and their respective segmentation masks. The data set is split into 70% training and 30% test data, resulting in 60 images for training and 15 for testing, which comparably is a very small data set to begin with. Other relevant information on preprocessing steps comprises the applied Z-score normalization,

[3] `https://github.com/MIC-DKFZ/nnUNet`

spacing (1.0, 1.0), and patch-size (768, 363), and, aiming to align pixel intensity, a consistent spacial scale across all images is applied, as well as memory efficiency for training [IF21, 205]. Further, cross validation (5-fold) is automatically executed to choose the best configuration, as stated in the supplementary material [IF21]. In addition, a combination of data augmentation techniques are applied (rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, gamma correction and mirroring) [IF21, 205].

- *Method* Information on applied methods can be found in the accompanying *plans.json* file when running the nnU-Net code. Once, the model configuration is chosen for application in the *main* version, this information possibly becomes relevant for other stakeholders at the interface with compliance.

- *Representation* This information is relevant to developers during model development, and possibly auditors at later stages.

- *Evaluation* Since non-architectural changes, such as preprocessing and post-processing, significantly impact model performance [IF21, 206], these automatic configurations could possibly be further refined during model optimization, if nnU-Net is selected.

8.2.1.2. QG nnU-Net

Closely linked to preprocessing is the model configuration. It comprises the combination of design decisions related to relevant hyperparameter during model training, as well as the model architecture, which is equally executed in an automatic manner when choosing nnU-Net. More information on possible design choices can be found in the supplementary material on nnU-Net design principles of the corresponding publication [IF21, 205].

- *Content* The nnU-Net v2 framework is utilized for model configuration. The *Residual Encoder UNet* is applied with a kernel size of (3, 3). During training, the model optimizes segmentation performance by minimizing the averaged Dice and CE loss. A polynomial learning rate schedule is applied, starting with an initial learning rate of 0.01. Optimization is performed using SGD with Nesterov momentum ($\mu = 0.99$) to enhance convergence stability and efficiency. The training process consists of 1.000 epochs, each comprising 250 minibatches, with foreground oversampling to improve class balance in highly imbalanced datasets. For inference, a sliding window approach with a half-patch size overlap is used, where Gaussian weighting at the patch center ensures smooth predictions and reduces edge artifacts. [IF21, 205]

Dice Score	0.8637
IoU	0.7719
Sensitivity	0.8307
Specificity	0.9723

Table 8.1.: nnU-Net [IF21] performance on the test set with only 60 labeled training images of the esophagus.

- *Method* The generated *plans.json* file provides information on the applied model architecture and related settings. In addition, the supplementary material offers a comprehensive overview on hyperparameter [IF21, 205].

- *Representation* This information is essential for developers during the model development process who applies the model architecture (and possibly, needs to be translated for auditors, as well).

- *Evaluation* If nnU-Net is selected, these automatic configurations could potentially be further refined during model optimization. However, the question remains if this is a reasonable direction for optimization, since the nnU-Net is designed to include automatic optimization tailored to the extracted data fingerprint.

8.2.1.3. QG Performance Metrics Compilation – Results

Based on the in the previously, in section 8.1.3.2 introduced performance evaluation metrics compilation, the nnU-Net is evaluated in addition to visual segmentation model output on the test set. The concrete results, as summarized in table 8.1, are appended to the *Content* dimension of the respective leaf-QG. In addition they are interpreted in the *Evaluation* information layer.

The results of the segmentation model indicate strong performance across key metrics, corresponding to the results as depicted in figure 8.5. The Dice Score/F1-Score of 0.8637 demonstrates a good balance between sensitivity and precision, reflecting accurate segmentation of the region of interest. The IoU of 0.7719 shows a solid overlap between predicted and ground truth areas, with the model performing well in handling over- and under-segmentation. Sensitivity (0.8307) indicates the model's ability to effectively detect the esophagus, while the high Specificity (0.9723) suggests that the model is effective at distinguishing background pixels, which is expected given the high imbalance between the background and the esophagus. Overall, these results point to a well-performing model with balanced detection capabilities, while fine-grained esophagus' shape segments need to be adjusted by the physician interacting with the model.

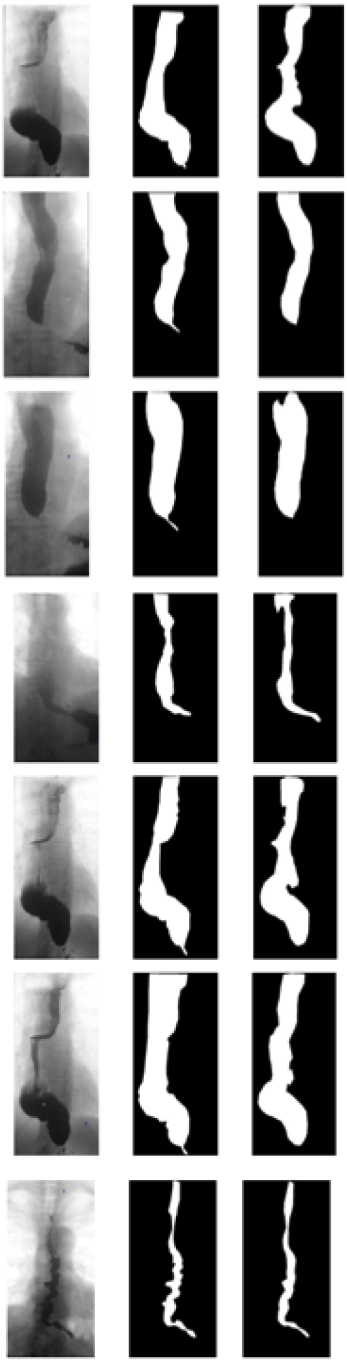


Figure 8.5.: Visualized segmentation performance of the nnU-Net on the test data with only 60 labeled training images. Raw input data (left) is depicted, in addition to the segmented mask (middle), and the segmentation model output (right). As apparent, the model achieves a good overall performance, while minor adjustments by the physician interacting with the model are required for an accurate shape.

8.2.2. v1b: Data Scarcity & Self-Supervised Learning (RCA-IUnet & Barlow-Twins)

The approach to compare the nnU-Net performance against is based on the BT-Unet framework [PA22b]. The authors introduce a self-supervised learning approach to enhance biomedical image segmentation, particularly in scenarios with limited annotated data, which is the case for our use case. By leveraging the Barlow Twins method [ZJ21a], BT-Unet pre-trains the encoder of a U-Net model through redundancy reduction, allowing it to learn meaningful data representations in an unsupervised manner. The full network is then fine-tuned for segmentation on the labeled data. Thus, this approach effectively utilizes both, annotated and unannotated samples. Among the tested U-Net architectures, *Residual Cross-spatial Attention guided Inception U-Net* (RCA-IUnet) [PA22c] demonstrates the highest performance gains, which is why this architecture was chosen as comparison for our simulation. [PA22b] However, as will become apparent in this section, it performs considerably worse than nnU-Net with our limited amount of labeled data and preprocessing optimization. Therefore, this approach is transferred to backlog for possible re-evaluation when more resources are available within our MQG4A simulation.

8.2.2.1. QG Data Preprocessing

Since the BT-Unet approach consists of two phases, i.e. self-supervised learning and fine-tuning, preprocessing is organized differently for unlabeled data, which is augmented in contrast to the labeled data.

- *Content* Generally, all images are resized to 256x256 pixels, and normalized by division with 255, resulting in all values ranging from 0 to 1. **self-supervised learning** was performed on 187 unlabeled TBE images on the entire data set. In addition, image augmentations are performed, resulting in two different images [PA22b, 4590]. In the context of our simulation, augmentation techniques include random resizing and cropping, horizontal flipping, rotation, each executed with 50% probability and solarization with 30% probability. For **fine-tuning**, the dataset, consisting of 85 labeled samples, is split into 70% training (60 samples) and 30% (15 samples) testing data. In addition, 5-fold cross-validation is applied.

- *Method* This approach is based on the BT-Unet framework [PA22b].

- *Representation* **Developers** need to access pre-processed data during model development (and possibly, auditors during compliance assessment).

- *Evaluation* This very rudimentary preprocessing could be further refined to achieve a better performing model. For instance, division by 255 for normalization could be replaced by Z-score normalization, as well as other settings could be adapted to the preprocessing, as carried out automatically by the nn-Unet framework to enhance comparability and possibly, performance. However, this is beyond the scope of our MQG4A simulation.

8.2.2.2. QG RCA-IUnet & Barlow Twins

This section outlines the BT-Unet approach based on the RCA-IUnet architecture [PA22c]. However, technical considerations are not the focus of our contribution, and for a more profound overview, it is advised to further consult the related publications. Further, the model code, as applied for the BT-Unet approach is open-source, and available on GitHub. [4]

- *Content* The **self-supervised learning** phase of the BT-Unet framework begins by passing an input image through the model's encoder, where a projection head processes the extracted feature maps into a compact representation. Once training is complete, the projection head is discarded, as its weights are not relevant for the actual segmentation task. The learned encoder weights are then transferred to the RCA-IUnet. [PA22b, 4591] "Adam optimizer with learning rate initialized at 1e−3 is used for all the experiments that decay by a factor of 0.1 once the learning stagnates for better segmentation results" [PA22b, 4592]. For **Fine-tuning**, 5-fold cross validation is applied, and the soft dice loss minimized, while for pre-training, the cross-correlation loss (Barlow twins) [ZJ21a] is minimized [PA22b, 4592].

- *Method* Since in [PA22b], the RCA-IUnet architecture displays superior performance to other U-Net based architectures regarding data scarcity in the experiments carried out by the authors, this architecture was selected to simulate model configuration intra-selection [PA22b, 4594].

- *Representation* The **Developer** needs to understand and implement the applied configuration (and possibly, auditors).

- *Evaluation* Possibly, other technical configurations perform better. For instance, with respect to performance comparison against nnU-Net, preprocessing steps could be aligned, as well as more TBE data annotated.

[4] `https://github.com/nspunn1993/RCA-IUnet/blob/main/models.py`

8.2.2.3. QG Performance Metrics Compilation – Results

RCA-IUnet with Barlow Twins is evaluated based on the, in the previous section 8.1.3.2, introduced performance evaluation metrics compilation. Additionally, visual segmentation model output on the test set provides insights into model performance. The concrete results, as summarized in Table 8.2, are appended to the *Content* dimension of the respective leaf-QG. In addition they are interpreted in the *Evaluation* information processing layer for further optimization measures.

Dice Score	0.3966
IoU	0.2593
Sensitivity	0.753
Specificity	0.4233

Table 8.2.: RCA-IUnet with Barlow Twins [PA22b] performance on the test set with only 60 labeled training images of the esophagus and limited preprocessing. Evidently, the model failed to learn the esophagus' shape and instead seems to having defaulted to conservatively predicting background pixels. This tendency is reflected in Figure 8.6.

The segmentation performance of RCA-IUnet with Barlow Twins in its current state is not convincing, as reflected by the evaluation metrics, and visually in Figure 8.6. The Dice Score (0.3966) and IoU (0.2593) indicate poor overlap between the predicted segmentation and ground truth, suggesting significant over- or undersegmentation. In this case probably over-segmentation prevails, which would explain the relatively high sensitivity (0.753). The model seems to struggle with precise boundary delineation, as it tends to classify large portions of the TBE image as the esophagus, without capturing its correct shape. Continuing this tendency, the model's specificity (0.4233) is notably low, meaning the model fails to distinguish background pixels effectively, probably resulting in more random that systematized outputs. Overall, these results indicate that the current amount of labeled data is insufficient for the model to learn a reliable segmentation representation, requiring further refinement or additional labeled samples to improve performance. Given the data imbalance between esophagus versus background pixels, a higher specificity compared to sensitivity is typically expected. However, the observed results suggest that the model architecture may be too complex for the current amount of labeled data. This aligns with previous experiments conducted by the authors, where data scarcity was still in the range of several hundred samples. [PA22b, 4592]. In addition, as previously stated, optimizing preprocessing steps could significantly enhance model performance. These considerations are documented within MQG4A for future work, while equally supporting a high AI literacy among project teams.

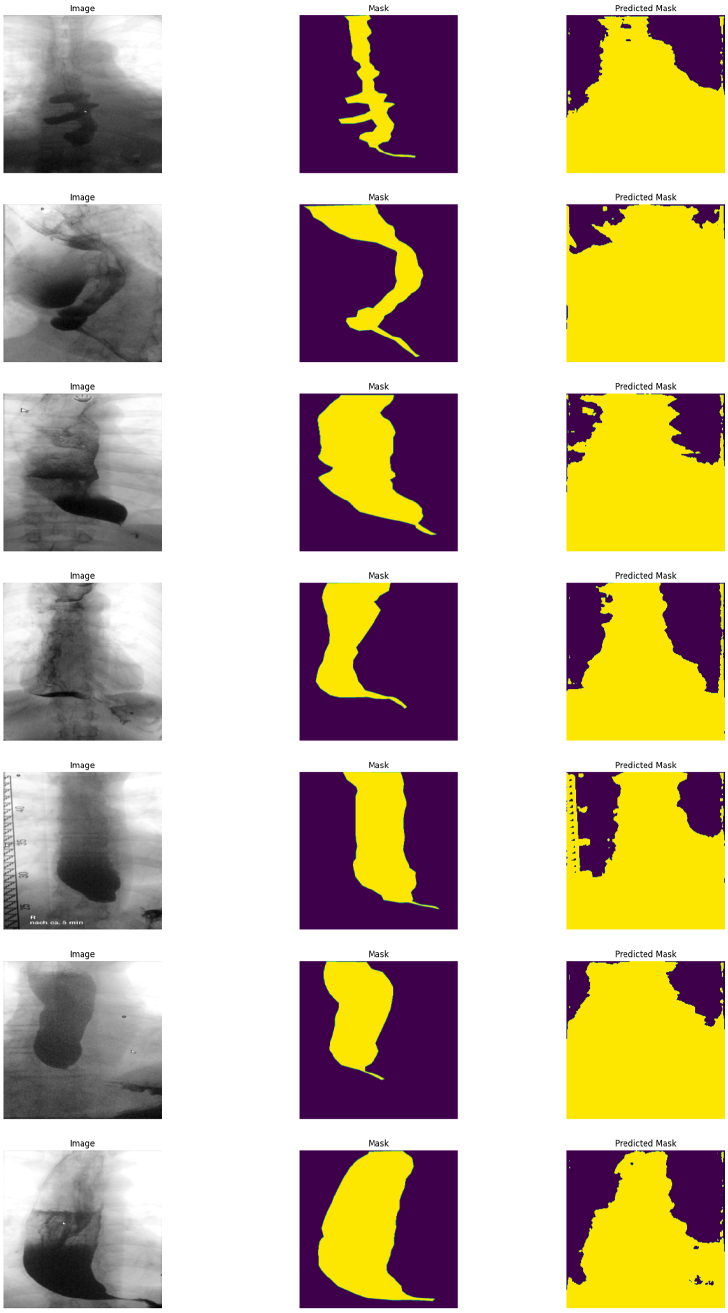


Figure 8.6.: Visualized segmentation performance of the RCA-IUnet using Barlow Twins on the test data with only 60 labeled training images and limited preprocessing. Raw input data (left) is depicted, in addition to the segmented mask (middle), and the segmentation model output (right). As apparent, the model seemingly did not learn anything about the esophagus' shape, and "plays safe" with predicting background pixels.

8.3. *Main*-Template (v2): Post-Selection & Outlook

After running the intra-selection experiments, reflecting the design decision which model configuration to select for TBE segmentation, the results clearly indicate that, with the current state of labeled data and preprocessing knowledge for medical image segmentation, nnU-Net is the preferable choice for the prototypical software *EsophagusVisualization*. As a result, the post-selection version (v3) appends the nnU-Net model configuration to the MQG4A lifecycle design, as introduced in section 8.1 for v1. This may include additional related information in form of QGs, such as raw model output transformations that are required for embedding the model within the software workflow. Aligning the design choice with the *main* lifecycle version equally comprises comprehensive information linking, as introduced in this section. In addition, the resulting QG compilation may include further design decisions. For instance, following our illustrated model development flow, this refers to methods to optimize the selected segmentation model configuration in form of e.g. hyperparameter tuning, optimized preprocessing, as well as more labeled data. We propose to underpin such design decision-making analogously to intra-selection (v2), in form of model snap-shots, that combine design choices with concrete results for decision-making. However, this needs to be evaluated based on more use cases, as will be analyzed in the next section 8.4, where we evaualte the demonstrated MQG4A simulation.

8.3.1. Information Linking

Information linking is crucial in designing the AI lifecycle, as it ensures seamless integration between different stages and design decisions, from data collection to model deployment and monitoring. Therefore, aiming to enhance traceability, reproducibility, transparency and compliance, the leaf-QG template provides information layers to extract interdependencies along the AI lifecycle, as introduced in section 4.3. The following sections illustrate information linking for the novel design choice on the TBE segmentation model configuration with the *main* lifecycle template.

8.3.1.1. QG Data Preprocessing

We outline information linking for data preprocessing in the following for our MQG4A simulation. This information is appended to the other leaf-QG dimensions, as previously introduced, in section 8.2.1.

- *Input Information* refers to required information to conduct data preprocessing. Consequently, knowledge on the **Data Design Input**, as introduced in v1 pre-selection, is required to understand the data format, based on which preprocessing is conducted. Further, data quality which is shaped by data collection, as well as insights obtained through data analysis, comprise relevant considerations for preprocessing [KM22]. Finally, information on the **Model Configuration** shapes preprocessing choices. Relevant system-related information comprises stakeholder, such as the **Developer**.

- *Output Information* Preprocessing prepares the data for the model, therefore, the **Model Configuration** shapes and is impacted by preprocessing steps. Automatizing the connection between input data, data preprocessing and model architecture can be interpreted as the key research question of the nnU-Net framework [IF21], and other approaches exist that aim to enhance this pipeline for medical images [MK21b] [MJ24]. These pipelines provide an alternative to manual **Optimization** steps.

- *Risk Management* RM related to data preprocessing is beyond the scope of our MQG4A simulation. For instance, preprocessing methods could be interpreted as a risk control mitigating inaccurate model output, highlighting the relation between accurate preprocessing and model quality [IF21] [VD23]. Vice versa, inaccurate per-processing methods pose additional risks related to model output quality.

8.3.1.2. QG nnU-Net

The following information linking is appended to the, previously introduced leaf-QG dimensions for nnU-Net model configuration.

- *Input Information* that shapes the model configuration is derived from **Data Design Input**, e.g. which type of model to apply, which in our case results in gray scale (TBE) image segmentation. Additionally, the model configuration and **data preprocessing** steps are closely linked [PB24], as previously outlined, resulting in bi-directional connections. Relevant system information comprises the active stakeholder **Developer** during the development phase.

- *Output Information* The model configuration comprises the model architecture, as well as hyperparameters, which can both be optimized during **Optimization**, which may include additional post-processing steps [AD22] [Fur21].

- *Risk Management* Within our MQG4A simulation for model selection, RM is centered around **inaccurate model output** and the human-in-the-loop, as introduced in section 8.1. An unfitting model architecture for the use case at hand results in inaccurate model output, while a fitting one mitigates this risk. Proving the model architecture is suitable, can be based on MQG4A intra-selection template versions that document the underlying selection logic. Towards further risk mitigation, it should be clearly documented when and how the model is optimized or retrained, e.g. based on a reasonable threshold regarding the amount of labeled data. This strategy may include revisiting previous intra-selection versions for retraining and reevaluation.

8.3.1.3. QG Performance Metrics Compilation – nnU-Net

The previously, in section 8.1 introduced *QG Performance Metrics Compilation* is extended with concrete results generated by nnU-Net, as detailed in section 8.2.

8.3.2. Resulting *Model Configuration* Template

This section provides an overview of the resulting information structure for TBE segmentation within *EsophagusVisualization* and summarizes the previously identified system and RM information blocks, as well as lifecycle QGs. It is showcased on GitHub.[5]

System Information Block As a first step after pulling the MQG4A template from MQG4DK, system-related information, which is accessible to relevant stakeholders is inserted. Within our simulation, this comprises the following list:

- The TBE segmentation model is defined, as embedded within *EsophagusVisualization*. This comprises information on its capabilities and intended use, environment, and degree of automation, among others.

- Domain Knowledge on Achalasia is provided.

- Documentation is simulated, focusing on the user manual, and model installation instructions.

- Stakeholders are included, highlighting the active developer, the consulting physician, and the passive physician who interacts with the model within the software.

[5] `https://github.com/miriamelia/MQG4AI/blob/main/README.md`

Risk Management In alignment with information on the system, we illustrate RM centered around *inaccurate model output* and the human-in-the-loop, resulting in the following compilation of related risks:

- **Inaccurate model output**
  Technical Robustness and Safety – Accuracy
- **Lack of collaboration mechanisms**
  Diversity, Non-discrimination, & Fairness – Stakeholder Participation
- **Waste of computing resources**
  Societal & Environmental Well-being – Environmental Well-being
- **Automation bias**
  Human Agency & Oversight – Human Oversight

Lifecycle Finally, centered around the design decision which model configuration to select, which resulted in choosing the nnU-Net framework [IF21], we introduce a combination of leaf-QGs:

- **Data Design Input**
  Data – Acquisition
- **Data preprocessing**
  Data – Utilization
- **Raw Model Output transformation**
  Data – Model Output
- **nnU-Net**
  Development – Model Configuration
- **Performance Metrics**
  Development – Model Evaluation
- **Feedback Loops**
  Maintenance – Support

8.4. Evaluation

In this chapter we simulate MQG4A template versions structured according to pre-, intra- and post-selection steps [EM24] for lifecycle planning, and documentation of design decisions, as well as related information, centered around *model configuration* for image segmentation. This scenario is based on a retrospective analysis of results, provided by the supervised bachelor's thesis *Oesophageal Segmentation using Barlow Twins and Multi-Label Annotation in the context of the Rare Disease Achalasia – A Self-Supervised Learning Approach*. Our focus lies on showcasing MQG4AI building blocks, and transferring the envisioned concept of MQG4A interaction, rather than the technical implementation, which functions as foundation to apply the lifecycle planning blueprint. This section evaluates our retrospective analysis, highlighting MQG4AI's customizable and generic character, that underlines the need to implement the MQG4AI software as a flexible lifecycle building kit. This aims to ensure sufficient customizability, while simultaneously defining standardized template application scenarios, as well as a generic lifecycle structure in alignment with MQG4DK.

Customizable MQG4AI Components MQG4AI incorporates a combination of components that can be applied to design the AI lifecycle in a flexible manner, addressing the iterative character of empirical design decision-making and lifecycle evolutions. Leaf-QGs, for instance, provide an information structure that outlines how information on individual design decisions can be structured. In our scenario, all preprocessing and model configuration steps are consolidated within a single leaf-QG respectively. This decision is based on the fact that the corresponding *Design Decisions* (e.g. which optimizer, loss function, or augmentation techniques to apply) were largely predetermined and, as explained in the leaf-QG method dimension, are grounded in existing publications [IF21] [PA22b]. Theoretically, the *MQG4Application* interaction allows for the creation of individual optimization template versions for each of these design decisions (or combinations thereof). However, the level of complexity required depends on how extensive a design decision is considered and whether empirical experiments are conducted to support the decision-making structure. For instance, common hyperparameter tuning approaches that extract optimal combinations of design decisions from a pool of possible values, such as grid or random search [IJ24], through multiple runs can be summarized as multiple *model configuration* leaf-QGs and related *performance evaluation* QGs, including concrete results within one intra-selection MQG4A template version for documentation, if deemed reasonable. Generally, MQG4AI provides customizability to organize design decisions in a fine-grained manner, or in a combined view. Another option to organize related design decisions in contrast to pre-selection template versions could be realized through appending leaf-QGs to leaf-QGs, resulting in a hierarchical structure within leaf-QG definition. This is possibly more suitable if design decisions are closely linked, which is the case for the data splitting strategy and pre-processing steps which needs to be executed afterwards, to avoid pitfalls [MF22]. In our simulation, pre-selection comprises separate, "stand-alone" design decisions compared to model selection, i.e. performance metrics selection and data design input. Overall, towards practical applicability of this approach, further evaluation is required.

Template Versioning Strategy In our example, we simulate a simple logic based on pre-, intra- and post-selection steps [EM24]: two segmentation model frameworks are tested, with each template version summarizing information related to the results. This generalizable design decision-making method conceptually aligns with pre-, in-, and post-processing methods towards bias mitigation that focus on the complete algorithm [SR22, 28] [NE20, 6], and is transferred to individual design decisions: data comprises the algorithm's input, analogously individual design decisions are built on previously established information. The algorithm, as well as any related design decision, is modifiable in itself. Finally, the algorithm's or design decision's results are adjustable. Alternative approaches to organize design decision-making are possible and the MQG4AI lifecycle blueprint should provide sufficient flexibility to allow for comprehensive design decision-making. The key aspect for choosing template versions-based design decision-making is that one version integrates design decisions corresponding to concrete results. This aims to document the lifecycle evolution while facilitating information distribution. In summary, the MQG4AI software needs to be developed in a manner that enables a customizable, interactive, and user-friendly application of the, in this work introduced MQG4AI components, analogously to a building kit that is in constant communication with MQG4DK, the living RAI lifecycle blueprint knowledge base.

MQG4DK Contributions Following DSR [vJ20], MQG4DK should continuously evolve. Potential MQG4A-based contributions within our simulation include metric combinations, and findings on the application of multi-label annotations, as well as analyzing their impact on model performance, once a sufficiently large labeled dataset becomes available. Due to the limited nature of our dataset, knowledge contributions about both model configurations within MQG4DK is currently difficult, apart from recognizing that nnU-Net [IF21] appears to be a promising option as state-of-the-art. It would be interesting to explore further optimizations, such as a combination of nnU-Net with Barlow Twins [ZJ21a] as future work. Additionally, further testing is required to determine whether RCA-IUnet [PA22c] within the Barlow Twins framework [PA22b] would perform better with an increased amount of labeled data and in-depth preprocessing phase design. Finally, the proposed supplementary information blocks on the intelligent system and related risks could provide valuable information for MQG4DK.

Open Questions Future research should primarily focus on the practical applicability of this scenario as an additional management layer for contributing stakeholders in individual projects. Due to the retrospective nature of our use case, this aspect is challenging to evaluate. Open questions include, for example, the substructure of the generic lifecycle, as described. Given the close interdependence between raw model output transformations, preprocessing steps, and model configuration, highlighting design decisions such as data splitting and cross-validation as training strategies, for instance, it might be more appropriate to integrate these as additional (sub-)leaf-QGs within model configuration, as previously outlined. Alternatively, if no complex, empirical evaluation strategy is associated with them, these design decisions could simply be included as informational components within the leaf-QG, summarizing model configuration, as with our simulation. MQG4AI's building blocks support such modularity, but further testing with real-world applications is necessary to identify standardizable component application strategies. All in all, our focus remains on the development stage, but evaluating MQG4AI's building blocks and their application is equally relevant for other stages such as deployment and maintenance. Moreover, leaf-QG tags, based on a ML design pattern structure [WH22] for MQG4A are not the focus of our evaluation. In MQG4DK, they serve the function of summarizing the information provided by individual leaf-QGs and enabling an intelligent search for MQG4A template-v0 configuration. However, their current structure appears suboptimal for practical applications. For instance, relevant considerations for an effective intelligent search within individual projects could include template versions and intra-version scenarios, which have been placed in the backlog for now, such as RCA-IUnet with Barlow Twins [PA22b], which might be re-evaluated with additional data at a later stage.

Ultimately, our simulation is merely a proposal intended to demonstrate the customizability of MQG4AI components. To enable comprehensive testing, it would be beneficial to implement these building blocks within an interactive and flexible software solution that allows users to independently create new QG instances or design tag structures. This may exclude push-to-MQG4DK operations, which may possibly result in the addition of two different layers of leaf-QG tags within the information processing template. Additionally, the envisioned tool should facilitate the seamless linking of bidirectional information, the creation of generalizable QG templates, and the integration of additional information layers. In summary, by enabling a flexible approach to shaping the AI lifecycle and its evolution, we aim to foster the implementation of RAI in a way that ensures transparency and adaptability.

Part IV.

CONCLUSION AND OUTLOOK

9 MQG4AI Roadmap

We investigate embedding quality and ethics with AI lifecycle design to support *Contributing Stakeholders*, while capturing technical dependencies and socio-ethical considerations, envisioning RAI systems. We subdivide this immense endeavor, which requires a broad community of diverse stakeholders, into four main challenges. Namely, we highlight lifecycle interdependenceis and AI pitfalls, diverse and distributed contextual information that needs to be considered, explore embedded ethics, as well as the evolving character of AI lifecycle design knowledge. In response, we introduce the generic and customizable MQG4AI lifecycle blueprint design kit, rooted in information and knowledge management. *Quality Gates* that construct the lifecycle in a flexible manner are at the heart of MQG4AI, emphasizing intra-lifecycle and contextual information linking, as well as guided design decision-making. *Information Blocks* summarize relevant supplementary design knowledge, illustrated for RM. The resulting dual system is based on principles from *Design Science Research* [vJ20], enabling public, ongoing, and decentralized knowledge updates (MQG4DK), as well as private, use case-specific application, highlighting collaborative design (MQG4A).

MQG4AI supports structured lifecycle planning, versioning, and interlinked IM in AI projects. In continuous alignment with a global, high-quality design knowledge base, the living lifecycle blueprint facilitates reliable and traceable design decisions across the AI lifecycle. MQG4AI is intended to support verification and validation procedures, so that "[...] the system [...] achieves its intended use in its intended operational environment" [ISO23c, 28]. It equally contributes to broader transparency dimensions, such as simulatability (how model behavior shifts with varying data), decomposability (understanding the model's individual components), and algorithmic clarity (how specific input combinations yield particular outputs) [Art24, 23]. Long-term, MQG4AI is envisioned to be implemented as a tool, emphasizing the generic and customizable character of its building blocks, as detailed throughout the previous chapters.

This chapter concludes our proposition. Section 9.1 first summarizes MQG4AI in relation to the research objectives outlined in section 1.3. Section 9.2 considers future work, building on the vision and limitations discussed in section 3.4, where we highlight considerations for implementing MQG4AI as a software.

9.1. MQG4AI Blueprint

MQG4AI emerges as a lifecycle blueprint, based on customizable building blocks to organize (implemented) RAI methods along the AI lifecycle, under the inclusion of relevant supplementary contextual information (oriented towards AI QM [Fut24]), and ethics, as detailed in part II. The resulting information structure can be viewed on GitHub[1]. While we aim for generalizability of the proposed setup, we focus on IM along the model development stage of high-risk applications in medicine, using DNNs. MQG4AI is to be comprehended as a *living blueprint*, that evolves in a decentralized manner with a growing number of application scenarios (MQG4A) and design contributions (MQG4DK), as outlined in section 3.3.7, drawing on principles from DSR [vJ20]. For this endeavor to be successful, the blueprint's design needs to be continued in a decentralized and on-going manner, benefiting from a critical mass of contributors. The following sections provide an overview of our proposition to design MQG4AI, aligned with our research objectives, before considering future work and MQG4AI as a tool, in the next section 9.2.

9.1.1. Designing the AI Lifecycle

Research Objective Provide structured guidance for reliable design decision-making and evaluation throughout the AI lifecycle by leveraging iterative, interdependent processes through generalizable and customizable building blocks. This approach enhances quality and promotes AI literacy, as detailed in sections 1.3.1.1, 1.3.2, and 1.3.3.

Focus We focus on the development stage for MQG4AI design, leveraging practical experience to identify a generic lifecycle flow, as introduced in section 4.2. Other lifecycle stages are explored more on a theoretical basis and only broached regarding relevant information extraction with respect to model development.

Multi-functional QGs comprise the core building block of MQG4AI. They provide a customizable structure for AI lifecycle-adjusted information planning, processing, and passing that is envisioned to support quality assurance towards RAI by design, as detailed in chapter 4. The concept *Quality Gate* is derived from traditional software QM and product development practices [Fil06] [Flo08] [GM09] [HS09], adapted to the AI context, and designed to incorporate AI-specific dynamics. To address AI's inherent complexity (e.g. its stochasticity and opacity), QGs mirror individual design decisions, and construct AI lifecycle stages in a hierarchical and dynamic manner, which results in generalizable, and system-specific QG graphs, highlighting interdependencies.

[1] `https://github.com/miriamelia/MQG4AI/blob/main/README.md`

We differentiate two general types of QGs for information structuring and processing: *Leaf-QGs* and *Collection-QGs*, that construct the AI lifecycle from generic process steps to use case-specific design decisions, that require an additional, and complex decision process. Together, they design the AI lifecycle in a generic and customizable manner, under consideration of interdependencies. Leaf-QGs, as detailed in section 4.3, are characterized by flexible information extraction layers, including relevant input- and output-information for design decision-making. For instance, system application-specific content, such as the *intended use and environment*, impact the implementation of specific lifecycle design decisions, which is made transparent within single leaf-QGs. In reverse, relevant information for e.g. the *post-market monitoring system* that needs to be documented [Fut24], is partly derived from technical lifecycle design choices, such as identified impact points of an evolving data distribution along the AI lifecycle. Such information can be collected based on reviewing all QGs combined. We propose to identify relevant information within the leaf-QG template, using the Socratic Method [IA12] to guide stakeholders through asking questions that support a systematic exploration of relevant considerations. Thus, we aim to close the gap to use case-specificity, and enhance AI literacy, long-term.

In addition, we propose a generic, high-level foundation of six stages to organize AI lifecycle design information, detailed in section 4.2 (*Conceptualization*, *Data*, *Development*, *Deployment*, *Maintenance*, *Decommissioning*). Their execution should be understood as iterative. Each lifecycle evolution is characterized by a distinct focus, depending on the project's progress, while all stages need to be addressed from the start of project *Conceptualization*. We aim to design the generic lifecycle in a way that ensures applicability across a wide range of use case-specific implementations, resulting in a modularizable structure that accommodates differing providers for specific lifecycle stages. For instance, the *Data* can be bought externally for in-house model *Development*, as well as particular *Deployment*, and *Maintenance* services, while sub-processes, such as *data preprocessing*, or the intelligent system's *on-boarding strategy* in its intended real world setting need to be developed in-house, in alignment with the specific use case.

In summary, QGs extract interdependencies in the form of a project-specific QG graph structure, featuring (bi-)directional vertical and horizontal information connections between various QGs, as well as linking supplementary, contextual RAI information. This approach results in merging conceptual information on AI projects, as exemplified in the next section, with a comprehensive AI system lifecycle in an organized manner along multiple sub-levels, aiming to guide reliable design and shed light on AI lifecycle evolutions from a conceptual view, which benefits overall transparency and AI literacy. This setup is intended to enable "[...] the participation of representatives from all sectors and types of organisations [...]" [SG24b, 2] in lifecycle design.

9.1.2. Dynamic RAI Information Blocks & Embedded Ethics

Research Objective Identify and link all information that is relevant to implement RAI systems, with the aim to allow for global AI design knowledge alignment, in the form of established formats, such as best practices and existing standards, towards quality and ethics by design, focusing on high-risk domains, as introduced in sections 1.3.1.2,1.3.1.3, and 1.3.3.

Focus We emphasize the integration of ethics within lifecycle design in the context of MQG4AI. Additionally, we concentrate on RM, including AI system-related contextual information, as a central component of AI QM [Fut24]. As a result, our proposed MQG4AI blueprint centers on three interconnected information blocks (*AI System*, *AI Lifecycle*, and *AI Risk Management*), with QGs embedded along the AI lifecycle. This customizable blueprint aligns with Article 17 of the AI Act [Fut24] and supports scalability through dynamic information linking, including communication with external RAI blocks and internal lifecycle connections via modular leaf-QG information processing layers, as described in section 3.4.1.2. Overall, this flexible IM approach enables system-wide evaluation and remains generalizable across domains, particularly for high-risk AI.

The proposed MQG4AI blueprint, as introduced in part II, is envisioned to provide a unifying approach towards implementing RAI through comprehensive IM, bridging implementation and the real world. Simply accessing code for compliance assessment does not suffice for intelligent systems and requires a different approach that considers "[...] the complexity of interpretation of code on execution (the models can change, self-learn and selfadapt) [...]" [GC20, 3]. This process unfolds in a dynamic, evolving real world context, where linking use case-specific contextual information to AI lifecycle implementations fosters RAI behavior. Generally, MQG4AI's design philosophy is based on four fundamental principles, aiming to include ethical topics from various perspectives towards the implementation of embedded ethics [MS20]: *Stakeholder Inclusion*, *Domain Embedding*, lifecycle *Interdependency Analysis*, emphasizing overall *Risk Mitigation*, as outlined in section 3.3.2. They are rooted in our expertise within the medical domain but intended to be generalizable, and this setup is envisioned to guide further design contributions in the same manner.

We design the exemplified AI system information block based on the *AI Risk Ontology* (AIRO), which incorporates AI Act terminology, and the ISO 31000 series of RM standards [GD22]. It is organized as four sub-modules, with the option to be extendable, following MQG4AI's customizable character. Section 3.3.3 introduces the AI system block in more detail. Within the context of our contribution, we concentrate on *Application*, which summarizes information on intended use and purpose, and the *Stakeholder* sub-unit, providing an overview of relevant roles. We propose to structure stakeholders into *active*, *contributing*, and *passive*, each of which require a different project-related lifecycle information access. Additionally, we append an *Ethics_Specific* and a *Domain Knowledge* section to *Application*, aiming to provide relevant, risk-mitigating information on the concrete application to all stakeholders, that interact with the lifecycle template in case of MQG4A. For a comprehensive integration of ethics along the AI lifecycle, we equally append an *Ethics_General* section to provide foundational knowledge on AI ethics. As a result, MQG4AI is designed to incorporate the interplay between knowledge on general ethical concepts and (shared) use case-specific ethical considerations within the AI system information block. We focus on the human influence along the AI lifecycle, which, if not addressed, poses risks due to rudimentary real world knowledge, variable AI literacy and an inherently biased human mind. The ethical foundation is envisioned to translate abstract ethical concepts for the individual AI use case through enabling a decentralized materials collection, with the aim to shape a RAI mindset for all contributing stakeholders in form of an applicable ethos. Our proposed ethical foundation, a holistic approach [EM25b], is detailed in section 3.2.2.3.

Addressing regulatory AI QM requirements, we exemplify the identification and collection of generalizable AI risks that are related with implemented lifecycle design decisions, aiming to align RM in a continuous manner. The proposed approach to manage risks is structured according to TAI criteria, as defined by the *Assessment List for Trustworthy AI* (ALTAI) [Hig20], emphasizing the on-going consideration of AI ethics. RM in the context of AI and MQG4AI is introduced in detail in sections 3.2 and 3.3.5. Individual risks are related with system-information and lifecycle QGs, allocating the required documentation through a transparent information flow. Overall, we aim for risk mitigation by design through connecting all relevant contextual and lifecycle implementation information, to conduct comprehensive and continuous RM, i.e. (residual) risk identification, evaluation, and control, a highly use case-specific process. These criteria align with Article 9 of the AI Act [Fut24], emphasizing the continuous character of RM. The required testing strategy is envisioned to be supported by different MQG4AI template versions (MQG4A), as introduced in the next section. As a result, RM comprises a fundamental component of continuous AI lifecycle planning within the MQG4AI blueprint.

9.1.3. Decentralized & Iterative Design

Research Objective Bridge generalizable AI guidelines with domain-specific implementations by incorporating evolving design knowledge throughout the AI lifecycle, fostering AI literacy and enabling informed decision-making across diverse use cases. Evaluate the envisioned approach on concrete use cases with diverse perspectives to extract generalizable workflows and provide customizability to different domains, as outlined in sections 1.3.1, 1.3.2, 1.3.3, and 1.3.4.

Focus We identified two interaction scenarios (MQG4DK & MQG4A) within MQG4AI. They are based on principles from DSR [vJ20], which is further outlined in section 3.3.7. Thus, we aim to enable decentralized and on-going design contributions, as well as application during individual projects. We evaluate these interaction scenarios from different angles, including two distinct medical use cases (see section 5), aiming to extract generalizable workflows that align with the identified RAI information collection.

DSR is centered around the continuous communication between an abstract design knowledge base, and concrete use cases to test/update/transform the accumulated and continuously evolving design knowledge, in our case on RAI lifecycle planning. Consequently, MQG4AI interaction results in two main perspectives *public MQG4DesignKnowledge* (MQG4DK), and *private MQG4Application* (MQG4A), that articulate the same high-level lifecycle information structure. MQG4A offers a lifecycle conceptualization template with shared access, including different template sub-versions for more complex design decision-making, while documenting the overall project-specific lifecycle evolution. The evolving templates are filled with concrete guidelines for individual AI lifecycle design based on the MQG4AI blueprint, including (future) use case-adapted MQG4DK (leaf-)QG contributions, that define (novel) AI design knowledge, possibly derived from MQG4A.

From the MQG4DK-viewpoint, the MQG4AI blueprint is intended to provide an information structure that enables summarizing the ever-evolving and dynamic RAI lifecycle design knowledge, including contextual information, highlighting AI risks. The growing design knowledge base can be interpreted as a *living RAI lifecycle blueprint*. Overall, we aim to enable the creation of a growing RAI ontology to organize generalizable instructions for the domain- and use case-adapted implementation. MQG4DK is envisioned to continuously grow in a decentralized manner, and we intend to design the foundation for further contributions based on the MQG4AI blueprint, as introduced in this thesis.

In chapter 6, we propose a generic design, and design workflow, for the lifecycle development model explanation stage, in cooperation with XAI experts [EM25a] and in alignment with the *IEEE Guide for an Architectural Framework for Explainable Artificial Intelligence* [Art24]. Additionally, we illustrate how to append design knowledge to MQG4DK in form of a technical guideline for explanation (SHAP, LIME) quality assessment [LGC24]. This workflow is extendable to design other, high-level lifecycle stages in more detail through the incorporation of existing best practices, or standards by design. Further, chapter 7 introduces a leaf-QG contribution to MQG4DK for reliable performance evaluation metrics focusing on multi-label ECG classification centered around a fictional medical use case, situated in EM. The guidelines are organized as a compilation of leaf-QGs along the AI lifecycle (focus development stage), and based on our previously published consideration on how to implement reliable performance evaluation metrics [EM24]. We emphsize design decision linking with AI system-specific information, as well as identified AI risks through leaf-QGs. The, in section 4.3.2 outlined *QG Naming* structure (*name*; *view*) is intended to organize design choice contributions according to e.g. shared structural similarities of AI techniques (i.e. proposed approach to define *view*) within MQG4DK.

MQG4A is envisioned to be applied as an additional management layer for AI information-based lifecycle conceptualization during individual AI projects [Eura], focusing on contributing stakeholders along the AI lifecycle. We envision the first MQG4A version (v0) to be pulled from MQG4DK. Via an intelligent tag search, as introduced in section 4.3.2, relevant design knowledge for the concrete use case is selected, if existent, and pulled, in addition to the generic MQG4AI blueprint structure. During project application, different MQG4A versions reflect the evolution of individual lifecycles for planning, and, in contrast to MQG4DK, they include concrete results. Analogously to Git branching[2], the subsequent template versions are envisioned to reflect the interplay between one *main* version that steadily grows as the project advances, and multiple *development* versions to test and document more intricate design decisions. For instance, focusing on model development, MQG4A versions are intended to structure the iterative process of reliable lifecycle design decision-making as part of a continuous *Conceptualization* stage.

This is exemplified in chapter 8 for segmentation model selection within the medical software *EsophagusVisualization*, comprising a human-in-the-loop approach. Based on a retro-spective analysis, we simulate template versions organized according to pre-, intra-, and post-selection functionalities for reliable lifecycle planning. Pre-, and post-versions summarize relevant input information, and integrate derived findings with the *main* version, while two distinct intra-selection *development* MQG4A template snapshots (RCA-IUnet [PA22c] vs. nnU-Net [IF21]) evaluate the fitting model configuration, that is eventually merged into the *main* version. Generally, when making a *Design Decision*, the process begins with identifying prerequisites, such as contextual information or data composition. Next, the design decision is implemented based on a fitting method, and possibly, related design decisions are identified to provide further clarity. Finally, optimization techniques are implemented to enhance the outcomes. This generic structure is extendable to establish relevant design choices during model development, such as reliable performance metrics. We aim to reflect this iterative strategy in the form of MQG4A versions to log results with tried out lifecycle concepts that, once sufficiently tested, result in the preferred design choice. Other generic lifecycle stages might need a different approach.

[2] `https://git-scm.com/book/en/v2/Git-Branching-Branches-in-a-Nutshell`

9.2. Future Work

By providing a generic and customizable approach to RAI lifecycle design, illustrated on GitHub[3], this work opens pathways for future research to deepen MQG4AI's operationalization, especially through software prototyping, stakeholder validation, and integration into AI QMS for further testing. This includes aligning design decisions with the proposed IM structure. Based on this thesis, we aim to evaluate whether others perceive MQG4AI as a meaningful contribution, and to develop workflows that enable the decentralized expansion of the proposed MQG4AI blueprint. Moving forward, refining the foundational structure to encompass additional risks, use cases, and lifecycle stages will depend on collaboration with a diverse pool of experts. The following summarizes our vision and limitations, which are detailed in section 3.4, and introduces directions for future work that are to be considered from both of MQG4AI's interaction views. MQG4A needs to be tested as an additional management layer within individual projects to support multi-stakeholder collaboration. In parallel, ongoing efforts should aim to further develop MQG4DK. This parallelization is feasible thanks to MQG4AI's generic, customizable building blocks. Our contribution to design the MQG4AI blueprint is not complete, and needs further refinement:

- *RAI Information* The adaptable and scalable setup needs to be tested for applicability, including the integration of additional, relevant *Information Blocks*. The generic design intends to support AI QMS [Fut24] requirements by enabling dynamic information linking to (external) RAI information blocks. Within the broader context of MQG4AI, risks are identified and organized based on TAI criteria [Hig20]. This approach aligns with emerging AI standards that "[...] should be aimed at identifying and mitigating risks of AI systems on the health, safety and fundamental rights of individuals. This is a novel aspect for AI standardisation, as the orientation of published and ongoing ISO/IEC work takes a very different approach in terms of risk objectives and definitions" [SG24b, 4]. However, our exemplified approach to RM may not be practical. This requires an evaluation of alternative AI RM schemes, and these considerations are closely linked to the AI system information block. Finally, methods to analyze the identified information flow that enable setting up a *QG Scoring System* need to be integrated, as introduced in section 4.1.3.

- *Generic AI Lifecycle* All lifecycle stages need to be assessed, including our attempt to contribute to designing a standardizable form of the development stage on lower levels. Ultimately all high-level lifecycle phases should be designed in more detail. Further, when identifying high-level collection-QGs, different lifecycle approaches need to be considered, including e.g.

[3] `https://github.com/miriamelia/MQG4AI/blob/main/README.md`

the utilization of automatized ML pipelines. Generally, application to various domain-specific contexts and design decisions supports the expanding coverage of all AI lifecycle stages and further establishes information fusion with supplementary contextual information.

- *Design Decision-Making* The proposed leaf-QG template information processing structure needs to be further evaluated, especially for compliance-related purpose fulfillment, and system-wide evaluation scenarios, as introduced in section 4.1. This includes the definition of optional vs. required layers, testing and further deepening the integration of the Socratic Method [IA12], as well as refining the QG information linking setup. QG tags and QG naming, as detailed in section 4.3.2, require practical evaluation and possibly need to be refined. *QG Relations*, such as QG inheritance, need to be further researched, possibly through the lens of OOP [Par20].

- *Use Case Categorization* Further research is needed to define an effective, and preferably adaptive classification of AI use cases and design choices. This would enable a refined QG naming and tagging approach, as well as support the organization of RAI design knowledge. Classification may rely on structural similarities across applications, as well as domain-specific requirements.

- *Generalizability & Adaptability* Future research should aim to enhance the generalizability of MQG4AI by addressing varying levels of design decision-making complexity. This includes evaluating how to effectively organize design decisions in terms of granularity and complexity, as well as refining template versioning strategies. To accommodate diverse development contexts, MQG4AI should continue to support alternative design decision-making approaches. While the building blocks are considered generic for DL-based systems, design variations may be needed for different domains, or general-purpose AI, with a broad spectrum of application scenarios. Deriving generalizable MQG4AI design workflows from these efforts will be essential for broader MQG4AI interaction.

In addition to refining the MQG4AI blueprint, and to foster application, future work should focus on implementing MQG4AI as a software. More profound considerations on architectural choices and user interfaces for the proposed dual system (MQG4DK and MQG4A) are outlined in section 3.4.2.2, highlighting parallization of future work in section 3.4.4. The following summarizes key guidelines for tool development, aiming to adapt to the evolving dynamics of AI and real world contexts. Overall, MQG4AI's building blocks should be designed in a flexible and customizable manner that enables on-going refinements and testing, incorporating DSR [vJ20]:

- *Flexible Building Blocks* QGs and information fusion with (existing) supplementary information blocks should be implemented in a flexible manner, that allows customizability of overall lifecycle design, including appending or removing (optional) information layers, design decisions, and lifecycle processes on lower levels. With time and more design knowledge contributions, MQG4AI is envisioned to provide solid and flexible design guidelines. On-going real world testing is thus enabled, to define standard component design and application strategies, possibly including dual-layer leaf-QG tag structures for MQG4DK and MQG4A.

- *Decentralized Setup* MQG4DK and MQG4A scenarios need to be considered for implementation. This includes the proposed intelligent search to pull configurable MQG4A versions using leaf-QGs, or determining how to sustain MQG4DK, which may require oversight by a designated responsible authority.

- *Bi-directional Information Linking* The tool should support users with connecting QGs to optimize usability and interdependency-handling. If a QG is referenced as input information, the respective QG should be automatically added within the output information layer of the corresponding QG.

Once, the fundamental MQG4AI blueprint is sufficiently outlined and QGs, including flexible *leaf-QG* information layers, as well as supplementary information blocks are implemented in a customizable manner, concrete MQG4A scenarios can be tested. Once, a setup to host the growing design knowledge base is developed, decentralized MQG4DK contributions are enabled for continuous MQG4AI interactions. This results in an on-going real world validation of the blueprint and associated workflows, including concrete MQG4A scenarios and sector-specific implementations. In summary, extensive further research and testing is needed on how to implement the generic AI lifecycle and its associated QGs. The proposed MQG4AI lifecycle building kit should be designed as a tool with flexibility in mind, allowing for the on-going creation, analysis, and comparison of various lifecycle concepts. These should equally include contextual information, oriented towards high-risk AI QM, that cover regulatory requirements, and embed ethics by design. In parallel, efforts should focus on continuing and refining the generic, compliance-oriented AI lifecycle blueprint, so that it is able to cover the multitude of existing AI application scenarios. To achieve this, AI use cases need to be organized, aiming to provide design guidelines as early and generalizable as possible. And hopefully, one day, the question how to design RAI lifecycles can be answered swiftly and reasonably semi-automated, contributing to stable and effortless AI governance.

Finally, as with our research contribution surrounding MQG4AI, we are already exploring new horizons. Continuing our collaboration [EM25a] with Alba M. López, Katherine Corredor, Prof. Dr. Albert Sabater, and Prof. Dr. Esteban Garcìa-Cuesta, we propose a *General Ethical AI Framework* that builds on this thesis, in combination with the ethical AI framework, as introduced in [GC20] that explores how to ethically guide AI system behavior in continuous interaction with the real world, from a technical angle.[4] Within that context, MQG4AI functions as a method to guide generic and customizable design decision-making that is directed towards the creation of AI, which behaves ethically within its intended real world setting, as described in [GC20]. We highlight the importance of domain embedding and the dynamic character of ethical considerations the AI may encounter. Further, future work will continue exploring human-AI collaboration. For instance, following an intriguing idea as proposed by Prof. Dr. Bernhard Bauer, we plan to explore the design of a first prototype, using GenAI. Other suitable formats include podcast production, and the extraction of shorter texts, such as articles, based on this thesis. With our contribution, we aim to inspire continued exploration and engagement from the broader AI community. Bridging the complexity of DL with multifaceted demands of real world applications presents considerable challenges and remains a vital area for future research.

The end is always a new beginning...

[4] `https://youtu.be/Ga2sW2N-PkU`

Part V.

Annex

Bibliography

[AC21] Asan, O. ; Choudhury, A.: Research Trends in Artificial Intelligence Applications in Human Factors Health Care: Mapping Review. In: *JMIR Hum Factors* 8 (2021), Jun, Nr. 2, e28236. http://dx.doi.org/10.2196/28236. – DOI 10.2196/28236. – ISSN 2292–9495

[AD22] Abeyrathna D., et a.: A Morphological Post-Processing Approach for Overlapped Segmentation of Bacterial Cell Images. In: *Machine Learning and Knowledge Extraction* 4 (2022), Nr. 4, 1024–1041. http://dx.doi.org/10.3390/make4040052. – DOI 10.3390/make4040052. – ISSN 2504–4990

[AEM18] Al-Emran M., et a.: The impact of knowledge management processes on information systems: A systematic review. In: *International Journal of Information Management* 43 (2018), 173-187. http://dx.doi.org/https://doi.org/10.1016/j.ijinfomgt.2018.08.001. – DOI https://doi.org/10.1016/j.ijinfomgt.2018.08.001. – ISSN 0268–4012

[AI21] Amir I., et a.: *SGD Generalizes Better Than GD (And Regularization Doesn't Help)*. https://arxiv.org/abs/2102.01117. Version: 2021

[AJ18] Adebayo J., et a.: Sanity Checks for Saliency Maps. In: Bengio S., et a. (Hrsg.): *Advances in Neural Information Processing Systems* Bd. 31, Curran Associates, Inc., 2018

[AM24a] Aboy M., et a.: Navigating the EU AI Act: implications for regulated digital medical products. In: *npj Digital Medicine* 7 (2024), Sep, Nr. 1, 237. http://dx.doi.org/10.1038/s41746-024-01232-3. – DOI 10.1038/s41746–024–01232–3. – ISSN 2398–6352

[AM24b] Adib, B.R. ; MD Kausik, A.K.: AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications. In: *Hybrid Advances* 7 (2024), 100277. http://dx.doi.org/https://doi.org/10.1016/j.hybadv.2024.100277. – DOI https://doi.org/10.1016/j.hybadv.2024.100277. – ISSN 2773–207X

[AMJ18] Alvarez-Melis, D. ; Jaakkola, T.S.: On the robustness of interpretability methods. In: *arXiv preprint arXiv:1806.08049* (2018). http://dx.doi.org/10.48550/arXiv.1806.08049. – DOI 10.48550/arXiv.1806.08049

[ARD09] Alonso-Ríos D., et a.: Usability: A Critical Analysis and a Taxonomy. In: *International Journal of Human–Computer Interaction* 26 (2009), Nr. 1, 53–74. http://dx.doi.org/10.1080/10447310903025552. – DOI 10.1080/10447310903025552

[Art24] ARTIFICIAL INTELLIGENCE STANDARDS COMMITTEE: IEEE Guide for an Architectural Framework for Explainable Artificial Intelligence. In: *IEEE Std 2894-2024* (2024), S. 1–55. http://dx.doi.org/10.1109/IEEESTD.2024.10659410. – DOI 10.1109/IEEESTD.2024.10659410

[AS19] AL AHBABI S., et a.: Employee perception of impact of knowledge management processes on public sector performance. In: *J. Knowl. Manag.* 23 (2019), S. 351–373. http://dx.doi.org/10.1108/JKM-08-2017-0348. – DOI 10.1108/JKM–08–2017–0348

[AS23] ALI S., et a.: Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. In: *Information Fusion* 99 (2023), 101805. http://dx.doi.org/https://doi.org/10.1016/j.inffus.2023.101805. – DOI https://doi.org/10.1016/j.inffus.2023.101805. – ISSN 1566–2535

[AZB23] AL-ZAITI, S. ; BOND, R.: Explainable-by-design: Challenges, pitfalls, and opportunities for the clinical adoption of AI-enabled ECG. In: *Journal of Electrocardiology* 81 (2023), 292-294. http://dx.doi.org/https://doi.org/10.1016/j.jelectrocard.2023.08.006. – DOI https://doi.org/10.1016/j.jelectrocard.2023.08.006. – ISSN 0022–0736

[BA19] BARREDO ARRIETA A., et a.: Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. In: *arXiv* (2019)

[BA20] BENJAMIN A.T., et a.: Deep learning for comprehensive ECG annotation. In: *Heart Rhythm* 17 (2020), Nr. 5, Part B, 881-888. http://dx.doi.org/https://doi.org/10.1016/j.hrthm.2020.02.015. – DOI https://doi.org/10.1016/j.hrthm.2020.02.015. – ISSN 1547–5271. – Digital Health Special Issue

[Bar24] BARASSI, V.: Toward a Theory of AI Errors: Making Sense of Hallucinations, Catastrophic Failures, and the Fallacy of Generative AI. In: *Harvard Data Science Review* (2024), nov 25, Nr. Special Issue 5. https://hdsr.mitpress.mit.edu/pub/1yo82mqa

[BB24] BHUYAN B.P., et a.: Neuro-symbolic artificial intelligence: a survey. In: *Neural Computing and Applications* 36 (2024), Jul, Nr. 21, 12809-12844. http://dx.doi.org/10.1007/s00521-024-09960-z. – DOI 10.1007/s00521–024–09960–z. – ISSN 1433–3058

[BC20] BANERJEE, D.N. ; CHANDA, S.S: *AI Failures: A Review of Underlying Issues*. https://arxiv.org/abs/2008.04073. Version: 2020

[BC24] BOVE C., et a.: *Why do explanations fail? A typology and discussion on failures in XAI*. https://arxiv.org/abs/2405.13474. Version: 2024

[BG99] BIDARD, C. ; GUIDO, E.: The concept of sector. In: *University of Antwerp, Faculty of Applied Economics, Working Papers* (1999), 01

[BG14] BOECKXSTAENS G.E., et a.: Achalasia. In: *Lancet* 383 (2014), S. 83–93

[BG20] BERTOSSI, L. ; GEERTS, F.: Data Quality and Explainable AI. In: *Journal of Data and Information Quality* 12 (2020), 04, S. 1–9. http://dx.doi.org/10.1145/3386687. – DOI 10.1145/3386687

[BH21] BRANKA H.M., et a.: *Explainable AI in Credit Risk Management*. 2021

[BH23] BOCHE H., et a.: Limitations of Deep Learning for Inverse Problems on Digital Hardware. In: *IEEE Trans. Inf. Theor.* 69 (2023), Dezember, Nr. 12, 7887–7908. http://dx.doi.org/10.1109/TIT.2023.3326879. – DOI 10.1109/TIT.2023.3326879. – ISSN 0018–9448

[BJ22] BERNER J., et a.: *The Modern Mathematics of Deep Learning*. http://dx.doi.org/{10.1017/9781009025096.002}. Version: dec 2022

[BJ23] BURIAK J.M., et a.: Best Practices for Using AI When Writing Scientific Manuscripts. In: *ACS Nano* 17 (2023), Mar, Nr. 5, 4091-4093. http://dx.doi.org/10.1021/acsnano.3c01544. – DOI 10.1021/acsnano.3c01544. – ISSN 1936–0851

[BL15] BLAGUS, R. ; LUSA, L.: Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models. In: *BMC Bioinformatics* 16 (2015), November, Nr. 1, S. 363

[BL24] BOULOGNE L.H., et a.: The STOIC2021 COVID-19 AI challenge: Applying reusable training methodologies to private data. In: *Medical Image Analysis* 97 (2024), 103230. http://dx.doi.org/https://doi.org/10.1016/j.media.2024.103230. – DOI https://doi.org/10.1016/j.media.2024.103230. – ISSN 1361–8415

[BMS21] BARJA-MARTINEZ S., et a.: Artificial intelligence techniques for enabling Big Data services in distribution networks: A review. In: *Renewable and Sustainable Energy Reviews* 150 (2021), 111459. http://dx.doi.org/https://doi.org/10.1016/j.rser.2021.111459. – DOI https://doi.org/10.1016/j.rser.2021.111459. – ISSN 1364–0321

[BN20] BOOTH, B.M. ; NARAYANAN, S.S.: Fifty Shades of Green: Towards a Robust Measure of Inter-annotator Agreement for Continuous Signals. In: *Proceedings of the 2020 International Conference on Multimodal Interaction*. New York, NY, USA : Association for Computing Machinery, 2020 (ICMI '20). – ISBN 9781450375818, 204–212

[BP20] BELLE, V. ; PAPANTONIS, I.: Principles and Practice of Explainable Machine Learning. In: *arXiv* (2020). https://arxiv.org/abs/2009.11698

[BR23] BOMMASANI R., et a.: The Foundation Model Transparency Index. In: *ArXiv* abs/2310.12941 (2023). https://api.semanticscholar.org/CorpusID:264306385

[BR25] BOMMASANI R., et a.: The 2024 Foundation Model Transparency Index. In: *arXiv* (2025). https://arxiv.org/abs/2407.12929

[Bra02] BRAF, E.: Knowledge or Information. In: LIU, et a. K. (Hrsg.): *Organizational Semiotics: Evolving a Science of Information Systems*. Boston, MA : Springer US, 2002. – ISBN 978–0–387–35611–2, 71–90

[Bre01] BREIMAN, L.: Statistical modeling: the two cultures. In: *Statist. Sci.* 16 (2001), Nr. 3, 199–231. http://dx.doi.org/10.1214/ss/1009213726. – DOI 10.1214/ss/1009213726. – ISSN 0883–4237. – With comments and a rejoinder by the author

[Bro95] BROOKE, J.: SUS: A quick and dirty usability scale. In: *Usability Eval. Ind.* 189 (1995), 11

[BS23] BAND S.S., et a.: Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. In: *Informatics in Medicine Unlocked* 40 (2023), 101286. http://dx.doi.org/https://doi.org/10.1016/j.imu.2023.101286. – DOI https://doi.org/10.1016/j.imu.2023.101286. – ISSN 2352–9148

[BX23] BAKER, S. ; XIANG, W.: *Explainable AI is Responsible AI: How Explainability Creates Trustworthy and Socially Responsible Artificial Intelligence*. https://arxiv.org/abs/2312.01555. Version: 2023

[CA24] CREMADES A., et a.: *Additive-feature-attribution methods: a review on explainable artificial intelligence for fluid dynamics and heat transfer*. https://arxiv.org/abs/2409.11992. Version: 2024

[Cha17] CHAUDRY, S.: *Achalasia*. https://www.youtube.com/watch?v=ElaLtb2GvYg. Version: 2017. – accessed April 2025

[CM25] CAIRE MJ., et a.: Physiology, Synapse. In: *StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing* (2025). https://www.ncbi.nlm.nih.gov/books/NBK526047/

[DA21] DATTA A., et a.: Machine Learning Explainability and Robustness: Connected at the Hip. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. New York, NY, USA : Association for Computing Machinery, 2021 (KDD '21). – ISBN 9781450383325, 4035–4036

[DA22] DE SILVA, D. ; ALAHAKOON, D.: An artificial intelligence life cycle: From conception to production. In: *Patterns* 3 (2022), Nr. 6, 100489. http://dx.doi.org/https://doi.org/10.1016/j.patter.2022.100489. – DOI https://doi.org/10.1016/j.patter.2022.100489. – ISSN 2666–3899

[Dep] DEPARTMENT OF GASTROENTEROLOGY, UNIVERSITY HOSPITAL AUGSBURG: *Achalasia*. https://www.uk-augsburg.de/einrichtungen/kliniken/iii-medizinische-klinik/schwerpunkte-und-leistungen/gastroenterologische-funktionsdiagnostik. – accessed April 25th 2025

[dH22] DE BRUIJN H., et a.: The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. In: *Government Information Quarterly* 39 (2022), Nr. 2, 101666. http://dx.doi.org/https://doi.org/10.1016/j.giq.2021.101666. – DOI https://doi.org/10.1016/j.giq.2021.101666. – ISSN 0740–624X

[DH23] DING H., et a.: The Application of Artificial Intelligence and Big Data in the Food Industry. In: *Foods* 12 (2023), Nr. 24. http://dx.doi.org/10.3390/foods12244511. – DOI 10.3390/foods12244511. – ISSN 2304–8158

[DJ06] DAVIS J., et a.: The Relationship between Precision-Recall and ROC Curves. In: *Proceedings of the 23rd International Conference on Machine Learning*. New York, NY, USA : Association for Computing Machinery, 2006 (ICML '06). – ISBN 1595933832, 233–240

[DJ09] DENG J., et a.: ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, S. 248–255

[DJ19] DEVLIN J., et a.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *arXiv* (2019). https://arxiv.org/abs/1810.04805

[DL22a] DECKERS, R. ; LAGO, P.: Systematic literature review of domain-oriented specification techniques. In: *Journal of Systems and Software* 192 (2022), 111415. http://dx.doi.org/https://doi.org/10.1016/j.jss.2022.111415. – DOI https://doi.org/10.1016/j.jss.2022.111415. – ISSN 0164–1212

[DL22b] DOMINGUES, N. ; LAGAREIRO, A.: Structure and development of a clinical decision support system: application in high Digestive Gastroenterology. In: *Anals of Clinical and Medical Case Reports* 8 (2022), 01. https://acmcasereport.org/wp-content/uploads/2023/06/ACMCR-v8-1730.pdf. – ISSN 2639–8109

[DRN23] DÍAZ-RODRÍGUEZ N., et a.: Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. In: *Information Fusion* 99 (2023), 101896. http://dx.doi.org/https://doi.org/10.1016/j.inffus.2023.101896. – DOI https://doi.org/10.1016/j.inffus.2023.101896. – ISSN 1566–2535

[DS21] DHANORKAR S., et a.: Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In: *Proceedings of the 2021 ACM Designing Interactive Systems Conference*. New York, NY, USA : Association for Computing Machinery, 2021 (DIS '21). – ISBN 9781450384766, 1591–1602

[DS22a] DESHPANDE, A. ; SHARP, H.: Responsible AI Systems: Who are the Stakeholders? In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA : Association for Computing Machinery, 2022 (AIES '22). – ISBN 9781450392471, 227–236

[DS22b] DUBEY S.R., et a.: Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark. In: *arXiv* (2022). https://arxiv.org/abs/2109.14545

[EB23] ELIA, M. ; BAUER, B.: A methodology based on quality gates for certifiable AI in medicine: towards a reliable application of metrics in machine learning. In: *ICSOFT* (2023). http://dx.doi.org/https://doi.org/10.5220/0012121300003538. – DOI https://doi.org/10.5220/0012121300003538

[EM23] ELIA M., et a.: Abstract: Precision Medicine for Achalasia Diagnosis: A Multi-modal and Interdisciplinary Aapproach for Training Data Generation. In: *20th IEEE International Symposium on Biomedical Imaging*. Colombia, Cartagena, 2023

[EM24] ELIA M., et a.: Towards Certifiable AI in Medicine: Illustrated for Multi-label ECG Classification Performance Metrics. In: *2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, 2024, S. 1–8

[EM25a] ELIA M., et a.: *MQG4AI Towards Responsible High-risk AI – Illustrated for Transparency Focusing on Explainability Techniques*. https://arxiv.org/abs/2502.11889. Version: 2025

[EM25b] ELIA M., et a.: Responsible AI, ethics, and the AI lifecycle: how to consider the human influence? In: *AI and Ethics* (2025), Mar. http://dx.doi.org/10.1007/s43681-025-00666-z. – DOI 10.1007/s43681–025–00666–z. – ISSN 2730–5961

[EM25c] ELIA M., et a.: *Supplementary Material to the Conference Paper "Towards Certifiable AI in Medicine: Illustrated for Multi-label ECG Classification Performance Metrics"*. http://dx.doi.org/10.5281/zenodo.14652465. Version: Januar 2025

[EM25d] ERIKSSON M., et a.: Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation. In: *arXiv* (2025). https://arxiv.org/abs/2502.06559

[ER24] ESHMAM R., et a.: Deep learning for medical image segmentation: State-of-the-art advancements and challenges. In: *Informatics in Medicine Unlocked* 47 (2024), 101504. http://dx.doi.org/https://doi.org/10.1016/j.imu.2024.101504. – DOI https://doi.org/10.1016/j.imu.2024.101504. – ISSN 2352–9148

[ES24] EVANS, H. ; SNEAD, D.: Why do errors arise in artificial intelligence diagnostic tools in histopathology and how can we minimize them? In: *Histopathology* 84 (2024), Nr. 2, 279-287. http://dx.doi.org/https://doi.org/10.1111/his.15071. – DOI https://doi.org/10.1111/his.15071

[Eura] EUROPEAN COMMISSION – AI OFFICE: *Risk management logic of the AI Act and related standards*. https://substack.com/redirect/ba283dbc-c00a-45b6-a048-b5ccac460be1?j=eyJ1IjoiMW9iZDRlIn0.o6qW6XeQkS6flUyMCdK4-YX6pyCMg3NbloCElNwnm0I, accessed June 1st 2024

[eurb] EUROSTAT, EUROPEAN UNION: *Cardiovascular diseases statistics*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Cardiovascular_diseases_statistics. – accessed January 27th 2025

[Eur17] EUROPEAN UNION: *Regulation (EU) 2017/745 of the European Parlia-*

ment and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance.). http://data.europa.eu/eli/reg/2017/745/oj. Version: May 2017

[Eur19] EUROPEAN COMMISSION AND DIRECTORATE-GENERAL FOR COMMUNICATIONS NETWORKS, CONTENT AND TECHNOLOGY: *Ethics guidelines for trustworthy AI*. Belgium, Brussels : Publications Office, 2019. http://dx.doi.org/doi/10.2759/346720. http://dx.doi.org/doi/10.2759/346720

[FF24] FRIEDRICH, S. ; FRIEDE, T.: On the role of benchmarking data sets and simulations in method comparison studies. In: *Biometrical Journal* 66 (2024), Nr. 1, 2200212. http://dx.doi.org/https://doi.org/10.1002/bimj.202200212. – DOI https://doi.org/10.1002/bimj.202200212

[FH22] FOIDL H., et a.: *Data Smells: Categories, Causes and Consequences, and Detection of Suspicious Data in AI-based Systems*. https://arxiv.org/abs/2203.10384. Version: 2022

[Fil06] FILHO, P.: Quality Gates in Use-Case Driven Development. In: *Proceedings of the 2006 International Workshop on Software Quality*. New York, NY, USA : Association for Computing Machinery, 2006 (WoSQ '06). – ISBN 1595933999, 33–38

[FL22] FLORIDI L., et a.: capAI: A procedure for conducting conformity assessment of AI systems in line with the EU Artifcial Intelligence Act. In: *SSRN* (2022), 03. http://dx.doi.org/http://dx.doi.org/10.2139/ssrn.4064091. – DOI http://dx.doi.org/10.2139/ssrn.4064091

[Flo08] FLOHR, T.: Defining Suitable Criteria for Quality Gates. In: *Software Process and Product Measurement*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2008. – ISBN 978–3–540–89403–2, S. 245–256

[FM20] FILZ M., et a.: Virtual Quality Gates in Manufacturing Systems: Framework, Implementation and Potential. In: *Journal of Manufacturing and Materials Processing* (2020). http://dx.doi.org/10.3390/jmmp4040106. – DOI 10.3390/jmmp4040106

[FS22] FRIEDRICH S., et a.: Regularization approaches in clinical biostatistics: A review of methods and their applications. In: *Stat Methods Med Res* 32 (2022), November, Nr. 2, S. 425–440

[Fuk21] FUKUSHIMA, K.: Artificial Vision by Deep CNN Neocognitron. In:

IEEE Transactions on Systems, Man, and Cybernetics: Systems 51 (2021), Nr. 1, S. 76–90. `http://dx.doi.org/10.1109/TSMC.2020.3042785`. – DOI 10.1109/TSMC.2020.3042785

[Fur21] Furtado, P.: Loss, post-processing and standard architecture improvements of liver deep learning segmentation from Computed Tomography and magnetic resonance. In: *Informatics in Medicine Unlocked* 24 (2021), 100585. `http://dx.doi.org/https://doi.org/10.1016/j.imu.2021.100585`. – DOI https://doi.org/10.1016/j.imu.2021.100585. – ISSN 2352–9148

[Fut24] Future of Life Institute (FLI): EU AI Act Explorer / FLI. Version: June 2024. `https://artificialintelligenceact.eu/ai-act-explorer/`. 2024. – Online. – accessed Nov 10th 2025

[FZ19] Fuzhen Z., et a.: A Comprehensive Survey on Transfer Learning. In: *CoRR* abs/1911.02685 (2019). `http://arxiv.org/abs/1911.02685`

[GA24a] Gour A., et a.: ECG Based Heart Disease Classification: Advancement and Review of Techniques. In: *Procedia Computer Science* 235 (2024), 1634-1648. `http://dx.doi.org/https://doi.org/10.1016/j.procs.2024.04.155`. – DOI https://doi.org/10.1016/j.procs.2024.04.155. – ISSN 1877–0509. – International Conference on Machine Learning and Data Engineering (ICMLDE 2023)

[GA24b] Grzybowski A., et a.: A History of Artificial Intelligence. In: *Clinics in Dermatology* 42 (2024), Nr. 3, 221-229. `http://dx.doi.org/https://doi.org/10.1016/j.clindermatol.2023.12.016`. – DOI https://doi.org/10.1016/j.clindermatol.2023.12.016. – ISSN 0738–081X. – Dermatology and Artificial Intelligence

[GC20] García-Cuesta, E.: *Artificial Intelligent Ethics in the Digital Era: an Engineering Ethical Framework Proposal*. `https://arxiv.org/abs/2002.07734`. Version: 2020

[GD22] Golpayegani D., et a.: AIRO: An Ontology for Representing AI Risks Based on the Proposed EU AI Act and ISO Risk Management Standards. In: *Volume 55: Towards a Knowledge-Aware AI*, 2022 (Studies on the Semantic Web). – ISBN 9781643683201, S. 51–65

[GGG20] Gómez-González, E. ; Gómez, E.: Artificial Intelligence in Medicine and Healthcare: applications, availability and societal impact / Publications Office of the European Union. Version: 2020. `http://dx.doi.org/10.2760/047666`. Luxembourg, Luxembourg, 2020. – Report

[GI16] GOODFELLOW I., et a.: *Deep Learning*. MIT Press, 2016. – http://www.deeplearningbook.org

[Gic23] GICHOYA, et a. J.W.: AI pitfalls and what not to do: mitigating bias in AI. In: *Br J Radiol* 96 (2023), September, Nr. 1150, S. 20230023

[Gil18] GILPIN, L.H.: Explaining Explanations: An Overview of Interpretability of Machine Learning. In: *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 2018, S. 80–89

[GM09] GIEBEL M., et a.: Improved innovation through the integration of quality gates into the enterprise and product lifecycle roadmaps. In: *Cirp Journal of Manufacturing Science and Technology* 1 (2009), S. 199–205. http://dx.doi.org/10.1016/J.CIRPJ.2008.10.004. – DOI 10.1016/J.CIRPJ.2008.10.004

[Gol24] GOLPAYEGANI, D. et a.: AI Cards: Towards an Applied Framework for Machine-Readable AI and Risk Documentation Inspired by the EU AI Act. In: JENSEN M., et a. (Hrsg.): *Privacy Technologies and Policy*. Cham : Springer Nature Switzerland, 2024. – ISBN 978–3–031–68024–3, S. 48–72

[GR18] GUIDOTTI R., et a.: A Survey of Methods for Explaining Black Box Models. In: *ACM Comput. Surv.* 51 (2018), August, Nr. 5. http://dx.doi.org/10.1145/3236009. – DOI 10.1145/3236009. – ISSN 0360–0300

[GS22] GUSTAFSSON S., et a.: Development and validation of deep learning ECG-based prediction of myocardial infarction in emergency department patients. In: *Scientific Reports* 12 (2022), 11. http://dx.doi.org/10.1038/s41598-022-24254-x. – DOI 10.1038/s41598–022–24254–x

[GV24] GRÜNHERZ V., et a.: Automatic three-dimensional reconstruction of the oesophagus in achalasia patients undergoing POEM: an innovative approach for evaluating treatment outcomes. In: *BMJ Open Gastroenterology* 11 (2024), Nr. 1. http://dx.doi.org/10.1136/bmjgast-2024-001396. – DOI 10.1136/bmjgast–2024–001396

[HA22] HAUSCHKE A., et a.: VDE SPEC 90012 V1.0 - VCIO based description of systems for AI trustworthiness characterisation. In: *VDE* (2022), 04

[HA23] HEDSTRÖM A., et a.: The Meta-Evaluation Problem in Explainable AI: Identifying Reliable Estimators with MetaQuantus. In: *Transactions on Machine Learning Research* (2023). https://openreview.net/forum?id=j3FK00HyfU. – ISSN 2835–8856

[HC21] HAOMIN C., et a.: INTRPRT: A Systematic Review of and Guidelines for

Designing and Validating Transparent AI in Medical Image Analysis. In: *CoRR* abs/2112.12596 (2021). https://arxiv.org/abs/2112.12596

[Hen22] Henry, et a. K.E.: Human–machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. In: *npj Digital Medicine* 5 (2022), Jul, Nr. 1, 97. http://dx.doi.org/10.1038/s41746-022-00597-7. – DOI 10.1038/s41746–022–00597–7. – ISSN 2398–6352

[Hig20] High-Level Expert Group on Artificial Intelligence (HLEG): The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment / European Commission, Directorate-General for Communications Networks Content & Technology. Version: 2020. http://dx.doi.org/doi/10.2759/002360. Publications Office, 2020. – Guidelines

[HK98] Hauser, J. ; Katz, G.: Metrics: you are what you measure! In: *European Management Journal* 16 (1998), Nr. 5, 517-528. http://dx.doi.org/https://doi.org/10.1016/S0263-2373(98)00029-2. – DOI https://doi.org/10.1016/S0263–2373(98)00029–2. – ISSN 0263–2373

[HL24] Heidemann L., et a.: The European Artificial Intelligence Act. In: *Fraunhofer IKS* (2024). http://dx.doi.org/10.24406/publica-3899. – DOI 10.24406/publica–3899

[HM21] Haakman M., et a.: AI lifecycle models need to be revised. In: *Empirical Software Engineering* 26 (2021), Jul, Nr. 5, 95. http://dx.doi.org/10.1007/s10664-021-09993-1. – DOI 10.1007/s10664–021–09993–1. – ISSN 1573–7616

[HM22] Huecker M., et a.: Emergency Medicine History and Expansion into the Future: A Narrative Review. In: *Western Journal of Emergency Medicine* 23 (2022), 04, S. 418–423. http://dx.doi.org/10.5811/westjem.2022.2.55108. – DOI 10.5811/westjem.2022.2.55108

[HS09] Hammers, C. ; Schmitt, R.: Governing the process chain of product development with an enhanced Quality Gate approach. In: *Cirp Journal of Manufacturing Science and Technology* 1 (2009), S. 206–211. http://dx.doi.org/10.1016/J.CIRPJ.2008.09.005. – DOI 10.1016/J.CIRPJ.2008.09.005

[HS20] Hong S., et a.: Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. In: *Proc. ACM Hum.-Comput. Interact.* 4 (2020), Mai, Nr. CSCW1. http://dx.doi.org/10.1145/3392878. – DOI 10.1145/3392878

[HSA22] HICKS S. A., et a.: On evaluation metrics for medical applications of artificial intelligence. In: *Scientific Reports* 12 (2022), Apr, Nr. 1, 5979. `http://dx.doi.org/10.1038/s41598-022-09954-8`. – DOI 10.1038/s41598–022–09954–8. – ISSN 2045–2322

[HW22a] HAVERKAMP W., et a.: EKG-Diagnostik mit Hilfe künstlicher Intelligenz: aktueller Stand und zukünftige Perspektiven – Teil 2. In: *Herzschrittmachertherapie + Elektrophysiologie* 33 (2022), September, Nr. 3, S. 305–311

[HW22b] HAVERKAMP W., et a.: EKG-Diagnostik mithilfe künstlicher Intelligenz: aktueller Stand und zukünftige Perspektiven – Teil 1. In: *Herzschrittmachertherapie + Elektrophysiologie* 33 (2022), Juni, Nr. 2, S. 232–240

[Hwa18] HWANG, T.: *Computational Power and the Social Impact of Artificial Intelligence*. `https://arxiv.org/abs/1803.08971`. Version: 2018

[HY23] HO Y.R., et a.: Thinking more wisely: using the Socratic method to develop critical thinking skills amongst healthcare students. In: *BMC Med Educ* 23 (2023), März, Nr. 1, S. 173

[IA12] ILKER, Y. ; ADEM, K.: What is Socratic Method?" The Analysis of Socratic Method through "Self Determination Theory" and "Unified Learning Model. In: *Procedia Computer Science* (2012), 01

[IEE22] IEEE Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence. In: *IEEE Std 2801-2022* (2022), S. 1–31. `http://dx.doi.org/10.1109/IEEESTD.2022.9812564`. – DOI 10.1109/IEEESTD.2022.9812564

[IF21] ISENSEE F., et a.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. In: *Nature Methods* 18 (2021), Feb, Nr. 2, 203-211. `http://dx.doi.org/10.1038/s41592-020-01008-z`. – DOI 10.1038/s41592–020–01008–z. – ISSN 1548–7105

[IJ24] ILEMOBAYO J., et a.: Hyperparameter Tuning in Machine Learning: A Comprehensive Review. In: *Journal of Engineering Research and Reports* 26 (2024), 06, S. 388–395. `http://dx.doi.org/10.9734/jerr/2024/v26i61188`. – DOI 10.9734/jerr/2024/v26i61188

[Ind24] INDYKOV, V.: Component-based Approach to Software Engineering of Machine Learning-enabled Systems. In: *2024 IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN)*, 2024, S. 250–252

[Ins23] INSTITUTE FOR QUALITY AND EFFICIENCY IN HEALTH CARE (IQWIG): *In brief: What is an electrocardiogram (ECG)? [Updated 2023 Jun 6]*. `https://www.ncbi.nlm.nih.gov/books/NBK536878/`. Version: 2023. – accessed January 27th 2025

[ISO16] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION: Medical devices — Quality management systems — Requirements for regulatory purposes - EN ISO 13485:2016. Geneva, CH, 2016. – Standard

[ISO19] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION: Medical devices - Application of risk management to medical devices - EN ISO 14971:2019. Geneva, CH, 2019. – Standard

[ISO21a] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION: Information Technology – Artificial Intelligence – Bias in AI Systems and Aided Decision Making – ISO/EC – TR 24027:2021 (E). Geneva, CH, 2021. – Standard

[ISO21b] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION: Health Informatice – Applications of machine learning technologies in imaging and other medical applications – ISO/EC – TR 24291:2021 (E). Geneva, CH, 2021. – Technical Report

[ISO22a] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION: Information technology — Artificial intelligence — Artificial intelligence concepts and terminology - ISO/IEC 22989:2022. Geneva, CH, 2022. – Standard

[ISO22b] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION: Information technology — Artificial intelligence — Assessment of machine learning classification performance. Geneva, CH, 2022. – Standard

[ISO23a] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION: Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems – ISO/IEC 25059. Geneva, CH, 2023. – Standard

[ISO23b] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION: Information technology — Artificial intelligence — Management system — ISO/IEC 42001:2023. Geneva, CH, 2023. – Standard

[ISO23c] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION: Information Technology – Artificial Intelligence – AI system lifecycle processes – ISO/IEC FDIS 5338:2023(E). Geneva, CH, 2023. – Standard

[Joh] JOHNER INSTITUTE: *ISO 14971 and Risk Management*. https://blog.johner-institute.com/category/iso-14971-risk-management/#:~:text=The%20ISO\2014971C0the\20standard,probability\20and%20severity%20of%20harm. accessed Feb 6th 2024

[JS23] JÖRG S., et a.: MedAIcine: A Pilot Project on the Social and Ethical Aspects of AI in Medical Imaging. In: STEPHANIDIS C.A., et a. (Hrsg.): *HCI International 2023 Posters*. Cham : Springer Nature Switzerland, 2023. – ISBN 978–3–031–35989–7, S. 455–462

[KA12] KRIZHEVSKY A., et a.: ImageNet Classification with Deep Convolutional Neural Networks. In: PEREIRA F., et a. (Hrsg.): *Advances in Neural Information Processing Systems* Bd. 25, Curran Associates, Inc., 2012

[KA21] KOSHIYAMA A., et a.: Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms. In: *SSRN Electronic Journal* (2021), 01. http://dx.doi.org/10.2139/ssrn.3778998. – DOI 10.2139/ssrn.3778998

[KA23] KASHOU A.H., et a.: ECG Interpretation Proficiency of Healthcare Professionals. In: *Current Problems in Cardiology* 48 (2023), Nr. 10, 101924. http://dx.doi.org/https://doi.org/10.1016/j.cpcardiol.2023.101924. – DOI https://doi.org/10.1016/j.cpcardiol.2023.101924. – ISSN 0146–2806

[KA24] KALE A.U., et a.: Detecting Algorithmic Errors and Patient Harms for AI-Enabled Medical Devices in Randomized Controlled Trials: Protocol for a Systematic Review. In: *JMIR Res Protoc* 13 (2024), Juni, S. e51614

[KB17] KINGMA, D.P. ; BA, J.: Adam: A Method for Stochastic Optimization. In: *arXiv* (2017). https://arxiv.org/abs/1412.6980

[KC19] KELLY C.J., et a.: Key challenges for delivering clinical impact with artificial intelligence. In: *BMC Medicine* 17 (2019), Oct, Nr. 1, 195. http://dx.doi.org/10.1186/s12916-019-1426-2. – DOI 10.1186/s12916–019–1426–2. – ISSN 1741–7015

[KJ24] KRUMME J., et a.: Never Mind the Codes of Conduct. DARE You to Tackle Ethics in Software Development for eHealth. In: *Studies in health technology and informatics* 316 (2024), 08, S. 2–6. http://dx.doi.org/10.3233/SHTI240330. – DOI 10.3233/SHTI240330. ISBN 9781643685335

[KM22] KIRAN M., et a.: A review: Data pre-processing and data augmentation techniques. In: *Global Transitions Proceedings* 3 (2022), Nr. 1, 91-99. http://dx.doi.org/https://doi.org/10.1016/j.gltp.2022.04.020. – DOI https://doi.org/10.1016/j.gltp.2022.04.020. – ISSN 2666–285X. – International Conference on Intelligent Engineering Approach(ICIEA-2022)

[KP15] KAHRILAS P.J., et a.: The Chicago Classification of esophageal motility disorders, v3.0. In: *Neurogastroenterology & Motility* 27 (2015), Nr. 2, 160-174. http://dx.doi.org/https://doi.org/10.1111/nmo.12477. – DOI https://doi.org/10.1111/nmo.12477

[KR25] KILIAN R., et a.: European AI Standards Technical Standardization and Implementation Challenges under the EU AI Act / German AI Association General Catalyst Institute. Version: 2025. https://ki-verband.de/wp-content/uploads/2025/03/Study_European-AI-Standards_FINAL_20250325.pdf. Germany, 2025. – Report

[KV22] KHARCHENKO V., et a.: Quality Models for Artificial Intelligence Systems: Characteristic-Based Approach, Development and Application. In: *Sensors (Basel, Switzerland)* 22 (2022). http://dx.doi.org/10.3390/s22134865. – DOI 10.3390/s22134865

[KW23] KHAN W., et a.: SQL and NoSQL Database Software Architecture Performance Analysis and Assessments—A Systematic Literature Review. In: *Big Data and Cognitive Computing* 7 (2023), Nr. 2. http://dx.doi.org/10.3390/bdcc7020097. – DOI 10.3390/bdcc7020097. – ISSN 2504–2289

[KY24] KUMAR Y., et a.: The AI-Powered Evolution of Big Data. In: *Applied Sciences* 14 (2024), Nr. 22. http://dx.doi.org/10.3390/app142210176. – DOI 10.3390/app142210176. – ISSN 2076–3417

[Leb23] LEBEN, D.: Explainable AI as evidence of fair decisions. In: *Front Psychol* 14 (2023), Februar, S. 1069426

[Lee21] LEE, K.J.: Chapter Seven - Architecture of neural processing unit for deep neural networks. Version: 2021. http://dx.doi.org/https://doi.org/10.1016/bs.adcom.2020.11.001. In: SHIHO, K. (Hrsg.) ; GANESH, C.D. (Hrsg.): *Hardware Accelerator Systems for Artificial Intelligence and Machine Learning* Bd. 122. Elsevier, 2021. – DOI https://doi.org/10.1016/bs.adcom.2020.11.001. – ISSN 0065–2458, 217-245

[LGC24] LÓPEZ, A.M. ; GARCÍA-CUESTA, E.: On the transferability of local model-agnostic explanations of machine learning models to unseen data. In: *2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS)*. Madrid, Spain, 2024, S. 1–10

[LH16] LAKKARAJU H., et a.: Interpretable Decision Sets: A Joint Framework for Description and Prediction. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. New York, NY, USA : Association for Computing Machinery, 2016, S. 1675–1684

[Lin18] LIN, et a. T.Y.: *Focal Loss for Dense Object Detection*. https://arxiv.org/abs/1708.02002. Version: 2018

[Lip16] LIPTON, Z.C.: The Mythos of Model Interpretability. In: *CoRR* abs/1606.03490 (2016). http://arxiv.org/abs/1606.03490

[LM20] LITVIŇUKOVÁ M., et a.: Cells of the adult human heart. In: *Nature* 588 (2020), September, Nr. 7838, S. 466–472

[LO17] LI O., et a.: *Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions*. https://arxiv.org/abs/1710.04806. Version: 2017

[LS19] LAPUSCHKIN S., et a.: Unmasking Clever Hans predictors and assessing what machines really learn. In: *Nature Communications* 10 (2019), Mar, Nr. 1. http://dx.doi.org/10.1038/s41467-019-08987-4. – DOI 10.1038/s41467–019–08987–4. – ISSN 2041–1723

[LS25] LEE S.U., et a.: *Responsible AI Question Bank: A Comprehensive Tool for AI Risk Assessment*. https://arxiv.org/abs/2408.11820. Version: 2025

[LSI17] LUNDBERG, S.M. ; SU-IN, L.: A Unified Approach to Interpreting Model Predictions. In: GUYON I., et a. (Hrsg.): *Advances in Neural Information Processing Systems* Bd. 30, Curran Associates, Inc., 2017

[LV20] LAKSHMANAN V., et a.: *Machine Learning Design Patterns: Solutions to Common Challenges in Data Preparation, Model Building, and MLOps*. O'Reilly Media, 2020 https://books.google.de/books?id=Y52uzQEACAAJ. – ISBN 9781098115784

[LW23] LEE W., et a.: AI Like ChatGPT, Users Like Us: How ChatGPT Drivers and AI Efficacy Affect Consumer Behaviour. In: *Virtual Economics* 6 (2023), 12. http://dx.doi.org/10.34021/ve.2023.06.04(3). – DOI 10.34021/ve.2023.06.04(3)

[LY21] LIU Y., et a.: Automatic Multi-Label ECG Classification with Category Imbalance and Cost-Sensitive Thresholding. In: *Biosensors* 11 (2021), Nr. 11. http://dx.doi.org/10.3390/bios11110453. – DOI 10.3390/bios11110453. – ISSN 2079–6374

[LY23] Lee Y., et a.: Computability of Optimizers. In: *arXiv* (2023). https://arxiv.org/abs/2301.06148

[MA15] Mosa A.S., et a.: Online electronic data capture and research data repository system for clinical and translational research. In: *Mo Med* 112 (2015), Januar, Nr. 1, S. 46–52

[MA16] Moonen A., et a.: Long-term results of the European achalasia trial: a multicentre randomised controlled trial comparing pneumatic dilation versus laparoscopic Heller myotomy. In: *Gut* 65:732-9 (2016)

[MA21a] Makris A., et a.: MongoDB Vs PostgreSQL: A comparative study on performance aspects. In: *GeoInformatica* 25 (2021), Apr, Nr. 2, 243-268. http://dx.doi.org/10.1007/s10707-020-00407-w. – DOI 10.1007/s10707–020–00407–w. – ISSN 1573–7624

[MA21b] Markus A.F., et a.: The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. In: *Journal of Biomedical Informatics* 113 (2021), 103655. http://dx.doi.org/https://doi.org/10.1016/j.jbi.2020.103655. – DOI https://doi.org/10.1016/j.jbi.2020.103655. – ISSN 1532–0464

[MA23] Mao A., et a.: *Cross-Entropy Loss Functions: Theoretical Analysis and Applications*. https://arxiv.org/abs/2304.07288. Version: 2023

[MA24] Metikoš, L. ; Ausloos, J.: The Right to an Explanation in Practice: Insights from Case Law for the GDPR and the AI Act. In: *Forthcoming in Law, Innovation, and Technology* 17 (2024), 8, Nr. 2. http://dx.doi.org/10.2139/ssrn.4996173

[MA25] Meinke A., et a.: *Frontier Models are Capable of In-context Scheming*. https://arxiv.org/abs/2412.04984. Version: 2025

[MC21] Martin C.H., et a.: Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. In: *Nature Communications* 12 (2021), Juli, Nr. 1, S. 4122

[MD22] Müller D., et a.: Towards a guideline for evaluation metrics in medical image segmentation. In: *BMC Research Notes* 15 (2022), Jun, Nr. 1, 210. http://dx.doi.org/10.1186/s13104-022-06096-y. – DOI 10.1186/s13104–022–06096–y. – ISSN 1756–0500

[MD23] Müller D., et a.: Towards Automated COVID-19 Presence and Severity Classification. In: *Caring is Sharing – Exploiting the Value in Data for*

Health and Innovation 302 (2023), S. 917–921. http://dx.doi.org/10.3233/SHTI230309. – DOI 10.3233/SHTI230309

[MF22] Maleki F., et a.: Generalizability of Machine Learning Models: Quantitative Evaluation of Three Methodological Pitfalls. In: *Radiol Artif Intell* 5 (2022), November, Nr. 1, S. e220028. http://dx.doi.org/10.1148/ryai.220028. – DOI 10.1148/ryai.220028

[MG17] Montavon G., et a.: Methods for Interpreting and Understanding Deep Neural Networks. In: *CoRR* abs/1706.07-979 (2017). http://arxiv.org/abs/1706.07979

[MH21] Muller H., et a.: The Ten Commandments of Ethical Medical AI. In: *Computer* 54 (2021), jul, Nr. 07, S. 119–123. http://dx.doi.org/10.1109/MC.2021.3074263. – DOI 10.1109/MC.2021.3074263. – ISSN 1558–0814

[MH23] Mojica-Hanke, A.: Towards machine learning guided by best practices. In: *arXiv* (2023). https://arxiv.org/abs/2305.00233

[MH24] Mintesnot H., et a.: Electrocardiography interpretation competency among pediatric and child health residents at Addis Ababa University, Ethiopia. In: *BMC Medical Education* 24 (2024), Dec, Nr. 1, 1548. http://dx.doi.org/10.1186/s12909-024-06614-5. – DOI 10.1186/s12909–024–06614–5. – ISSN 1472–6920

[Mit97] Mitchell, T.: *Machine Learning*. McGraw Hill, 1997 https://www.cs.cmu.edu/~tom/files/MachineLearningTomMitchell.pdf

[MJ24] Ma J., et a.: Segment anything in medical images. In: *Nature Communications* 15 (2024), Jan, Nr. 1, 654. http://dx.doi.org/10.1038/s41467-024-44824-z. – DOI 10.1038/s41467–024–44824–z. – ISSN 2041–1723

[MK21a] Macfarlane, P.W. ; Kennedy, J.: Automated ECG Interpretation—A Brief History from High Expectations to Deepest Networks. In: *Hearts* 2 (2021), Nr. 4, 433–448. http://dx.doi.org/10.3390/hearts2040034. – DOI 10.3390/hearts2040034. – ISSN 2673–3846

[MK21b] Müller, D. ; Kramer, F.: MIScnn: a framework for medical image segmentation with convolutional neural networks and deep learning. In: *BMC Medical Imaging* 21 (2021), Jan, Nr. 1, 12. http://dx.doi.org/10.1186/s12880-020-00543-7. – DOI 10.1186/s12880–020–00543–7. – ISSN 1471–2342

[MK24] Muyskens K., et a.: When can we Kick (Some) Humans "Out of the

Loop"? An Examination of the use of AI in Medical Imaging for Lumbar Spinal Stenosis. In: *Asian Bioethics Review* (2024), 05. http://dx.doi.org/10.1007/s41649-024-00290-9. – DOI 10.1007/s41649–024–00290–9

[ML24] MERCOLLI L., et a.: Towards quality management of artificial intelligence systems for medical applications. In: *Z Med Phys* 34 (2024), Februar, Nr. 2, S. 343–352

[MM24a] MEINIKHEIM M., et a.: Effect of AI on performance of endoscopists to detect Barrett neoplasia: A Randomized Tandem Trial. In: *Endoscopy* 0 (2024), 03. http://dx.doi.org/10.1055/a-2296-5696. – DOI 10.1055/a–2296–5696

[MM24b] MUJTABA, D.F. ; MAHAPATRA, N.R.: Fairness in AI-Driven Recruitment: Challenges, Metrics, Methods, and Future Directions. In: *arXiv* (2024). https://arxiv.org/abs/2405.19699

[MMH22] MAARTMANN-MOE H., et a.: Design decision competence: Supporting user participation in design decisions. In: *Proceedings of the Participatory Design Conference 2022 - Volume 2*. New York, NY, USA : Association for Computing Machinery, 2022 (PDC '22). – ISBN 9781450396813, 196–202

[MN21] MINH N.V., et a.: c-Eval: A Unified Metric to Evaluate Feature-based Explanations via Perturbation. In: *2021 IEEE International Conference on Big Data (Big Data)*, 2021, S. 927–937

[Mol25] MOLNAR, C.: *Interpretable Machine Learning*. 3. 2025 https://christophm.github.io/interpretable-ml-book. – ISBN 978–3–911578–03–5

[MRM24] MUSTROPH, H. ; RINDERLE-MA, S.: Design of a Quality Management System based on the EU Artificial Intelligence Act. In: *arxiv* (2024), 11. http://dx.doi.org/10.48550/arXiv.2408.04689. – DOI 10.48550/arXiv.2408.04689

[MS20] MCLENNAN S., et a.: An embedded ethics approach for AI development. In: *Nature Machine Intelligence* 2 (2020), 07, S. 1–3. http://dx.doi.org/10.1038/s42256-020-0214-1. – DOI 10.1038/s42256–020–0214–1

[MS23] MARTINEZ S., et a.: Current and Future Use of Artificial Intelligence in Electrocardiography. In: *Journal of Cardiovascular Development and Disease* 10 (2023), 04, S. 175. http://dx.doi.org/10.3390/jcdd10040175. – DOI 10.3390/jcdd10040175

[MS24] Malik S., et a.: Artificial intelligence and industrial applications-A revolution in modern industries. In: *Ain Shams Engineering Journal* 15 (2024), 06. http://dx.doi.org/10.1016/j.asej.2024.102886. – DOI 10.1016/j.asej.2024.102886

[Nata] National Cancer Institute SEER Training Modules, U.S. Department of Health and Human Services - National Institutes of Health: *Structure of the Heart*. https://www.ncbi.nlm.nih.gov/books/NBK536878/. – January 27th 2025

[Natb] National Institute of Neurological Disorders and Stroke, National Institutes of Health: *Brain Basics: Know Your Brain*. https://www.ninds.nih.gov/health-information/public-education/brain-basics/brain-basics-know-your-brain. – March 19th 2025

[Nat23] National Artificial Intelligence Strategy for Ireland (NSAI): TOP TEAM ON STANDARDS IN AI AI Standards & Assurance Roadmap Action under 'AI – Here for Good,' / National Standards Authority of Ireland (NSAI). Version: 2023. https://www.nsai.ie/. Ireland, 2023. – Publication

[NE20] Ntoutsi E., et a.: Bias in data-driven artificial intelligence systems—An introductory survey. In: *WIREs Data Mining and Knowledge Discovery* 10 (2020), 02, Nr. 3. http://dx.doi.org/https://doi.org/10.1002/widm.1356. – DOI https://doi.org/10.1002/widm.1356

[Nic00] Nickols, F.: The knowledge in knowledge management. In: *The Knowledge Management Yearbook 2000-2001* (2000), 01, S. 12–21

[OJ24] Oviedo J., et a.: ISO/IEC quality standards for AI engineering. In: *Computer Science Review* 54 (2024), 100681. http://dx.doi.org/https://doi.org/10.1016/j.cosrev.2024.100681. – DOI https://doi.org/10.1016/j.cosrev.2024.100681. – ISSN 1574–0137

[Org04] Organisation for Economic Co-Operation and Development (OECD) Statistics Canada: Measuring Knowledge Management in the Business Sector: First Steps, Knowledge management / OECD Publishing. Version: 2004. http://dx.doi.org/https://doi.org/10.1787/9789264100282-en. Paris, France, 2004. – Report

[Org23] Organisation for Economic Co-operation and Development (OECD): Common guideposts to promote interoperability in AI risk management. In: *OECD Artificial Intelligence Papers* No. 5 (2023), 11.

http://dx.doi.org/https://doi.org/10.1787/ba602d18-en. – DOI https://doi.org/10.1787/ba602d18–en

[Ost23] Ostherr, K.: The future of translational medical humanities: bridging the data/narrative divide. In: *Medical Humanities* 49 (2023), Nr. 4, 529–536. http://dx.doi.org/10.1136/medhum-2023-012627. – DOI 10.1136/medhum–2023–012627. – ISSN 1468–215X

[PA22a] Pfob A., et a.: Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison. In: *BMC Med Res Methodol* 22 (2022), November, Nr. 1, S. 282

[PA22b] Punn, N.S. ; Agarwal, S.: BT-Unet: A self-supervised learning framework for biomedical image segmentation using barlow twins with U-net models. In: *Machine Learning* 111 (2022), Dec, Nr. 12, 4585-4600. http://dx.doi.org/10.1007/s10994-022-06219-3. – DOI 10.1007/s10994–022–06219–3. – ISSN 1573–0565

[PA22c] Punn, N.S. ; Agarwal, S.: RCA-IUnet: a residual cross-spatial attention-guided inception U-Net model for tumor segmentation in breast ultrasound imaging. In: *Machine Vision and Applications* 33 (2022), Feb, Nr. 2, 27. http://dx.doi.org/10.1007/s00138-022-01280-3. – DOI 10.1007/s00138–022–01280–3. – ISSN 1432–1769

[Pal24] Palmiotto, F.: When Is a Decision Automated? A Taxonomy for a Fundamental Rights Analysis. In: *German Law Journal* 25 (2024), Nr. 2, S. 210–236. http://dx.doi.org/10.1017/glj.2023.112. – DOI 10.1017/glj.2023.112

[Par20] Parsons, D.: Objects Working Together: Association, Aggregation and Composition. In: *Foundational Java* (2020), 09, S. 141–175. http://dx.doi.org/10.1007/978-3-030-54518-5_7. – DOI 10.1007/978–3–030–54518–5_7. ISBN 978–3–030–54517–8

[PB21] Paulsen B.A., et a.: Functional Safety Concept EGAS for Medical Devices. In: *Current Directions in Biomedical Engineering* 7 (2021), Nr. 2, 739–742. http://dx.doi.org/doi:10.1515/cdbme-2021-2189, Abruf: 2024-02-06. – DOI doi:10.1515/cdbme–2021–2189

[PB24] Provencher B., et a.: Hyperparameter tuning for deep learning semantic image segmentation of micro computed tomography scanned fiber-reinforced composites. In: *Tomography of Materials and Structures* 5 (2024), 100032. http://dx.

doi.org/https://doi.org/10.1016/j.tmater.2024.100032. – DOI https://doi.org/10.1016/j.tmater.2024.100032. – ISSN 2949–673X

[PD20] PETERS D., et a.: Responsible AI—Two Frameworks for Ethical Design Practice. In: *IEEE Transactions on Technology and Society* 1 (2020), Nr. 1, S. 34–47. http://dx.doi.org/10.1109/TTS.2020.2974991. – DOI 10.1109/TTS.2020.2974991

[PJ21a] PEREIRA J., et a.: High-accuracy protein structure prediction in CASP14. In: *Proteins: Structure, Function, and Bioinformatics* 89 (2021), Nr. 12, 1687-1699. http://dx.doi.org/https://doi.org/10.1002/prot.26171. – DOI https://doi.org/10.1002/prot.26171

[PJ21b] PEREIRA J., et a.: High-accuracy protein structure prediction in CASP14. In: *Proteins: Structure, Function, and Bioinformatics* 89 (2021), 07. http://dx.doi.org/10.1002/prot.26171. – DOI 10.1002/prot.26171

[PT23] PRINZ T., et a.: Market access of continuous learning AI systems in medicine. In: *VDE DGBMT* (2023)

[QJ23] QI J., et a.: *The Art of SOCRATIC QUESTIONING: Recursive Thinking with Large Language Models*. https://arxiv.org/abs/2305.14999. Version: 2023

[QL24] QIONG L., et a.: Operationalizing AI in Future Networks: A Bird's Eye View from the System Perspective. In: *arXiv* (2024). https://arxiv.org/abs/2303.04073

[Qui86] QUINLAN, J.R.: Induction of decision trees. In: *Machine Learning* 1 (1986), Mar, Nr. 1, 81–106. http://dx.doi.org/10.1007/BF00116251. – DOI 10.1007/BF00116251. – ISSN 1573–0565. – visited on 2019-09-11

[RA20] RIBEIRO A.H., et a.: Automatic diagnosis of the 12-lead ECG using a deep neural network. In: *Nature Communications* 11 (2020), apr, Nr. 1. http://dx.doi.org/10.1038/s41467-020-15432-4. – DOI 10.1038/s41467–020–15432–4

[RA21] REINKE A., et a.: *Common Limitations of Image Processing Metrics: A Picture Story*. http://dx.doi.org/10.48550/ARXIV.2104.05642. Version: 2021

[RA23] RAD A., et a.: Questionnaire "Artificial Intelligence (AI) in medical devices" / German association of notified bodies (IG-NB). 2023. – Guidelines

[RA24] REUEL A., et a.: *BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices*. https://arxiv.org/abs/2411.12990. Version: 2024

[RC22] RÖMMELE C., et a.: An artificial intelligence algorithm is highly accurate for detecting endoscopic features of eosinophilic esophagitis. In: *Scientific Reports* 12 (2022), 07. http://dx.doi.org/10.1038/s41598-022-14605-z. – DOI 10.1038/s41598–022–14605–z

[RG09] REMENYI, D. ; GRIFFITHS, P.: The Socratic Dialogue in the Work Place: Theory and Practice. In: *Electronic Journal of Knowledge Maangement* Vol.7 No.1 (2009), 01

[RH22] REIS, J. ; HOUSLEY, M.: *Fundamentals of Data Engineering*. O'Reilly Media, 2022 https://www.oreilly.com/library/view/fundamentals-of-data/9781098108298/. – ISBN 9781098108304

[Rib16] RIBEIRO, et a. M.T.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *CoRR* abs/1602.04938 (2016). http://arxiv.org/abs/1602.04938

[RK23] RAHMANI K., et a.: Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. In: *International Journal of Medical Informatics* 173 (2023), 104930. http://dx.doi.org/https://doi.org/10.1016/j.ijmedinf.2022.104930. – DOI https://doi.org/10.1016/j.ijmedinf.2022.104930. – ISSN 1386–5056

[RM21a] REYNA M.A., et a.: Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021. In: *2021 Computing in Cardiology (CinC)*, 2021, S. 1–4

[RM21b] ROBERTS M., et a.: Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. In: *Nature Machine Intelligence* 3 (2021), Mar, Nr. 3, 199-217. http://dx.doi.org/10.1038/s42256-021-00307-0. – DOI 10.1038/s42256–021–00307–0. – ISSN 2522–5839

[RM22] REYNA M.A., et a.: Issues in the automated classification of multilead ecgs using heterogeneous labels and populations. In: *Physiol Meas* 43 (2022), August, Nr. 8

[RM24] RYAN M.L., et a.: Integrating Artificial Intelligence Into the Visualization and Modeling of Three-Dimensional Anatomy in Pediatric Surgical Patients. In: *Journal of Pediatric Surgery* 59 (2024), Nr. 12, S. 161629. http://dx.doi.org/https://doi.org/10.1016/j.jpedsurg.

2024.07.014. – DOI https://doi.org/10.1016/j.jpedsurg.2024.07.014. – ISSN 0022–3468

[RN24] ROGER N., et a.: Architecting ML-enabled systems: Challenges, best practices, and design decisions. In: *Journal of Systems and Software* 207 (2024), 111860. http://dx.doi.org/https://doi.org/10.1016/j.jss.2023.111860. – DOI https://doi.org/10.1016/j.jss.2023.111860. – ISSN 0164–1212

[RS21] REDDY S., et a.: Evaluation framework to guide implementation of AI systems into healthcare settings. In: *BMJ health & care informatics* 28 (2021), 10. http://dx.doi.org/10.1136/bmjhci-2021-100444. – DOI 10.1136/bmjhci–2021–100444

[RSB18] ROBNIK-SIKONJA, M. ; BOHANEC, M.: Perturbation-Based Explanations of Prediction Models. In: *Springer International Publishing* (2018), 159–175. https://doi.org/10.1007/978-3-319-90403-0_9. ISBN 978–3–319–90403–0

[Rud17] RUDER, R.: An overview of gradient descent optimization algorithms. In: *arXiv* (2017). https://arxiv.org/abs/1609.04747

[Rus25] RUSCHEMEIER, H.: Thinking Outside the Box? In: STEFFEN, Bernhard (Hrsg.): *Bridging the Gap Between AI and Reality*. Cham : Springer Nature Switzerland, 2025. – ISBN 978–3–031–73741–1, S. 318–332

[SA20] SENIOR A.W., et a.: Improved protein structure prediction using potentials from deep learning. In: *Nature* 577 (2020), Jan, Nr. 7792, 706-710. http://dx.doi.org/10.1038/s41586-019-1923-7. – DOI 10.1038/s41586–019–1923–7. – ISSN 1476–4687

[San20] SANTHANAM, P.: Quality Management of Machine Learning Systems. In: SHEHORY O., et a. (Hrsg.): *Engineering Dependable and Secure Machine Learning Systems*. Cham : Springer International Publishing, 2020. – ISBN 978–3–030–62144–5, S. 1–13

[Sch19] SCHMIDT, R.M.: Recurrent Neural Networks (RNNs): A gentle Introduction and Overview. In: *arXiv* (2019). https://arxiv.org/abs/1912.05911

[SD86] SCHON, D.A. ; DESANCTIS, V.: The Reflective Practitioner: How Professionals Think in Action. In: *The Journal of Continuing Higher Education* 34 (1986), Nr. 3, 29–30. http://dx.doi.org/10.1080/07377366.1986.10401080. – DOI 10.1080/07377366.1986.10401080

[SD20] SHANAHAN, J.G. ; DAI, L.: Introduction to Computer Vision and Real Time Deep Learning-based Object Detection. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, NY, USA : Association for Computing Machinery, 2020 (KDD '20). – ISBN 9781450379984, 3523–3524

[SD24] SASANKA SEKHAR, C. ; DEBARAG NARAYAN, B.: Omission and commission errors underlying AI failures. In: *AI & SOCIETY* 39 (2024), Jun, Nr. 3, 937-960. http://dx.doi.org/10.1007/s00146-022-01585-x. – DOI 10.1007/s00146–022–01585–x. – ISSN 1435–5655

[SE10] STEYERBERG E.W., et a.: Assessing the performance of prediction models: a framework for traditional and novel measures. In: *Epidemiology* 21 (2010), Nr. 1, S. 128–38. http://dx.doi.org/10.1097/EDE.0b013e3181c30fb2. – DOI 10.1097/EDE.0b013e3181c30fb2. – ISSN 1044–3983. – Keywords: Epidemiologic Studies; Models, Statistical; Prognosis; ROC Curve; Reproducibility of Results; Risk Assessment

[SF19] SOKOL, K. ; FLACH, P.A.: Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. In: *CoRR* abs/1912.05100 (2019). http://arxiv.org/abs/1912.05100

[SF21] STIELER F., et a.: Towards Domain-Specific Explainable AI: Model Interpretation of a Skin Image Classifier Using a Human Approach. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021, S. 1802–1809

[SF23] STIELER F., et a.: LIFEDATA - A Framework for Traceable Active Learning Projects. In: *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, 2023, S. 465–474

[SG13] SAMHITA, L. ; GROSS, H.J.: The "Clever Hans Phenomenon" revisited. In: *Communicative & integrative biology* 6 (2013), Nov, Nr. 6, e27122-e27122. http://dx.doi.org/10.4161/cib.27122. – DOI 10.4161/cib.27122. – ISSN 1942–0889. – 24563716[pmid]

[SG24a] SASTRY G., et a.: *Computing Power and the Governance of Artificial Intelligence*. https://arxiv.org/abs/2402.08797. Version: 2024

[SG24b] SOLER G.J., et a.: Harmonised Standards for the European AI Act / European Commission. Version: 2024. Seville (Spain) : European Commission, 2024. – Forschungsbericht

[SH07] SOYER H.P., et a.: *Color Atlas of Melanocytic Lesions of the Skin*.

Springer Berlin Heidelberg, 2007 https://books.google.de/books?id=qN63CWMnRpUC. – ISBN 9783540351061

[SM08] SÜDHOF, T. C. ; MALENKA, R. C.: Understanding synapses: past, present, and future. In: *Neuron* 60 (2008), November, Nr. 3, S. 469–476

[SM09] SOKOLOVA M., et a.: A systematic analysis of performance measures for classification tasks. In: *Information Processing & Management* 45 (2009), Nr. 4, 427-437. http://dx.doi.org/https://doi.org/10.1016/j.ipm.2009.03.002. – DOI https://doi.org/10.1016/j.ipm.2009.03.002. – ISSN 0306–4573

[SM23] SICA M., et a.: 3D Model Artificial Intelligence-Guided Automatic Augmented Reality Images during Robotic Partial Nephrectomy. In: *Diagnostics* 13 (2023), 11, Nr. 22, S. 3454. http://dx.doi.org/10.3390/diagnostics13223454. – DOI 10.3390/diagnostics13223454

[SN14] SRIVASTAVA N., et a.: Dropout: a simple way to prevent neural networks from overfitting. In: *J. Mach. Learn. Res.* 15 (2014), Januar, Nr. 1, S. 1929–1958. – ISSN 1532–4435

[SN20] STRODTHOFF N., et a.: Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL. In: *CoRR* abs/2004.13701 (2020). https://arxiv.org/abs/2004.13701

[SP09] SCHWARTZ P.J., et a.: Prevalence of the congenital long-QT syndrome. In: *Circulation* 120 (2009), Oktober, Nr. 18, S. 1761–1767

[SQ21] SONG Q.C., et a.: Making Sense of Model Generalizability: A Tutorial on Cross-Validation in R and Shiny. In: *Advances in Methods and Practices in Psychological Science* 4 (2021), Nr. 1. http://dx.doi.org/10.1177/2515245920947067. – DOI 10.1177/2515245920947067

[SR22] SCHWARTZ R., et a.: *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. http://dx.doi.org/https://doi.org/10.6028/NIST.SP.1270. Version: 2022-03-15 04:03:00 2022

[SS24] SAIFULLAH S., et a.: The privacy-explainability trade-off: unraveling the impacts of differential privacy and federated learning on attribution methods. In: *Frontiers in Artificial Intelligence* 7 (2024), Juli. http://dx.doi.org/10.3389/frai.2024.1236947. – DOI 10.3389/frai.2024.1236947. – ISSN 2624–8212

[ST15] SAITO T., et a.: The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.

In: *PLOS ONE* 10 (2015), 03, Nr. 3, 1-21. http://dx.doi.org/10.1371/journal.pone.0118432. – DOI 10.1371/journal.pone.0118432

[Süd18] SÜDHOF, T. C.: Towards an Understanding of Synapse Formation. In: *Neuron* 100 (2018), Oktober, Nr. 2, S. 276–293

[SZ21] SUSSKIND Z., et a.: *Neuro-Symbolic AI: An Emerging Class of AI Workloads and their Characterization*. https://arxiv.org/abs/2109.06133. Version: 2021

[Tab23] TABASSI, E.: Artificial Intelligence Risk Management Framework (AI RMF 1.0) / National Institute of Standards and Technology, U.S. Department of Commerce. Version: 2023. http://dx.doi.org/https://doi.org/10.6028/NIST.AI.100-1. Washington, D.C., 2023. – Framework

[TJ21] TOHKA J., et a.: Evaluation of machine learning algorithms for health and wellness applications: A tutorial. In: *Computers in Biology and Medicine* 132 (2021), 104324. http://dx.doi.org/https://doi.org/10.1016/j.compbiomed.2021.104324. – DOI https://doi.org/10.1016/j.compbiomed.2021.104324. – ISSN 0010–4825

[TJ24] TERVEN J., et a.: *Loss Functions and Metrics in Deep Learning*. https://arxiv.org/abs/2307.02694. Version: 2024

[VA17] VASWANI A., et a.: Attention Is All You Need. In: *CoRR* abs/1706.03762 (2017). http://arxiv.org/abs/1706.03762

[van20] VANVROONHOVEN, J.: Medical Device White Paper Series. Risk management for medical devices and the new BS EN ISO 14971 / BSI (National Standards Body) Standards Ltd. 2020. – White Paper

[van21] VAN DER VEER, et a. S.N.: Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. In: *Journal of the American Medical Informatics Association* 28 (2021), 08, Nr. 10, 2128-2138. http://dx.doi.org/10.1093/jamia/ocab127. – DOI 10.1093/jamia/ocab127. – ISSN 1527–974X

[VD23] VALENKOVA D., et a.: Data Preprocessing Influence on the Medical Images Segmentation Quality. In: *2023 Seminar on Information Computing and Processing (ICP)*, 2023, S. 154–157

[vJ20] VOM BROCKE J., et a.: Introduction to Design Science Research. In: VOM BROCKE J., et a. (Hrsg.): *Design Science Research. Cases*, Springer, February 2020 (Progress in IS), S. 1–13

[VL20] VILONE, G. ; LONGO, L.: Explainable Artificial Intelligence: a Systematic Review. In: *arXiv* (2020)

[WC22] WIMMY C., et a.: Legal Analysis. European legislative proposal draft AI act and MDR/IVDR / Axon science based lawyers. Version: 2022. https://hooghiemstra-en-partners.nl/report-ai-act-in-relation-to-mdr-and-ivdr/?lang=en. 2022. – Forschungsbericht

[WH22] WASHIZAKI H., et a.: Software-Engineering Design Patterns for Machine Learning Applications. In: *Computer* 55 (2022), Nr. 3, S. 30–39. http://dx.doi.org/10.1109/MC.2021.3137227. – DOI 10.1109/MC.2021.3137227

[WJ24] WEISZ J.D., et a.: Design Principles for Generative AI Applications. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ACM, Mai 2024, 1–22

[Wol55] WOLFE, M.: The Concept of Economic Sectors. In: *The Quarterly Journal of Economics* 69 (1955), Nr. 3, 402–420. http://www.jstor.org/stable/1885848. – ISSN 00335533, 15314650

[Wor] WORLD HEALTH ORGANIZATION (WHO): *Cardiovascular diseases (CVDs)*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). – accessed January 27th 2025

[WS20] WANG S.Y., et a.: Big data requirements for artificial intelligence. In: *Curr Opin Ophthalmol* 31 (2020), September, Nr. 5, S. 318–323

[WX16] WU X., et a.: A Unified View of Multi-Label Performance Measures. (2016). http://dx.doi.org/10.48550/ARXIV.1609.00288. – DOI 10.48550/ARXIV.1609.00288

[WY19] WERNER Y.B., et a.: Endoscopic or Surgical Myotomy in Patients with Idiopathic Achalasiay. In: *N Engl J Med* 381:2219-2229 (2019)

[XX20] XU X., et a.: Interpretation of Electrocardiogram (ECG) Rhythm by Combined CNN and BiLSTM. In: *IEEE Access* 8 (2020), S. 125380–125388. http://dx.doi.org/10.1109/ACCESS.2020.3006707. – DOI 10.1109/ACCESS.2020.3006707

[YF19] YANG F., et a.: Evaluating Explanation Without Ground Truth in Interpretable Machine Learning. In: *CoRR* abs/1907.06831 (2019). http://arxiv.org/abs/1907.06831

[Yok19] YOKOYAMA, H.: Machine Learning System Architectural Pattern for Improving Operational Stability. In: *2019 IEEE International Conference on Software Architecture Companion (ICSA-C)*, 2019, S. 267–274

[ZA25] ZABOLI A., et a.: Exploring ChatGPT's potential in ECG interpretation and outcome prediction in emergency department. In: *The American Journal of Emergency Medicine* 88 (2025), 7-11. http://dx.doi.org/https://doi.org/10.1016/j.ajem.2024.11.023. – DOI https://doi.org/10.1016/j.ajem.2024.11.023. – ISSN 0735–6757

[ZJ21a] ZBONTAR J., et a.: Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In: MEILA, M. (Hrsg.) ; ZHANG, T. (Hrsg.): *Proceedings of the 38th International Conference on Machine Learning* Bd. 139, PMLR, 18–24 Jul 2021 (Proceedings of Machine Learning Research), 12310–12320

[ZJ21b] ZHOU J., et a.: Evaluating the quality of machine learning explanations: A survey on methods and metrics. In: *Electronics* 10 (2021), Nr. 5, S. 593. http://dx.doi.org/10.3390/electronics10050593. – DOI 10.3390/electronics10050593

[ZJ22] ZACHARIAS J., et a.: Designing a feature selection method based on explainable artificial intelligence. In: *Electronic Markets* 32 (2022), Dec, Nr. 4, 2159-2184. http://dx.doi.org/10.1007/s12525-022-00608-1. – DOI 10.1007/s12525–022–00608–1. – ISSN 1422–8890

[ZP25a] ZIETHMANN P., et a.: CDSS – An Interdisciplinary Perspective on the Statement of the Central Ethics Commission of the German Medical Association. In: *The Royal College of Radiologists Open* 3 (2025), 100163. http://dx.doi.org/https://doi.org/10.1016/j.rcro.2024.100163. – DOI https://doi.org/10.1016/j.rcro.2024.100163. – ISSN 2773–0662. – Global AI Conference 2025

[ZP25b] ZIETHMANN P., et a.: Clinical Decision Support Systems at the Intersection of Technology and Ethics: A Critical Analysis of the Ethical Guidelines Issued by the German Medical Association. In: *Digital Society* 4 (2025), Mar, Nr. 1, 15. http://dx.doi.org/10.1007/s44206-025-00175-w. – DOI 10.1007/s44206–025–00175–w. – ISSN 2731–4669

[ZV22] ZINCHENKO V.V., et a.: Methodology for Conducting Post-Marketing Surveillance of Software as a Medical Device Based on Artificial Intelligence Technologies. In: *Sovrem Tekhnologii Med* 14 (2022), September, Nr. 5, S. 15–23

[ZY24] ZHANG Y., et a.: Exploring the Application of the Artificial-

Intelligence-Integrated Platform 3D Slicer in Medical Imaging Education. In: *Diagnostics* 14 (2024), 01, S. 146. http://dx.doi.org/10.3390/diagnostics14020146. – DOI 10.3390/diagnostics14020146

[ZZ14] ZHANG, M. ; ZHOU, Z.: A Review on Multi-Label Learning Algorithms. In: *IEEE Transactions on Knowledge and Data Engineering* 26 (2014), Nr. 8, S. 1819–1837. http://dx.doi.org/10.1109/TKDE.2013.39. – DOI 10.1109/TKDE.2013.39

[ZZ21] ZHU Z., et a.: Identification of 27 abnormalities from multi-lead ECG signals: an ensembled SE_ResNet framework with Sign Loss function. In: *Physiol Meas* 42 (2021), Juni, Nr. 6

Glossary

AI Act The European legislation on AI aims to ensure fundamental rights, health, and safety. The AI Act classifies intelligent systems into four levels of risk: minimal, limited-, high-, and unacceptable risk, depending on their impact on the real world, with augmenting requirements, until prohibition. [Fut24]. 3, 76

AI lifecycle All design decisions and processes that comprise the transition from idea (and research) of intelligent systems, to production. High-level phases are generalizable across use cases, and we propose six generic phases in this thesis (*Conceptualization*, *Data*, *Development*, *Maintenance*, *Deployment*, and *Decommissioning*) that are executed in an iterative manner. Beyond the provider-perspective, the AI lifecycle closes the gap to compliance assessment. 4, 65, 76, 79, 168

AI Literacy Sufficient level of knowledge on how to address AI, tailored to the project, stakeholder role and type of interaction with the intelligent system. It comprises "[...] skills, knowledge and understanding that allow providers, deployers and affected persons, taking into account their respective rights and obligations in the context of this Regulation, to make an informed deployment of AI systems, as well as to gain awareness about the opportunities and risks of AI and possible harm it can cause", as described in Article 3 of the AI Act [Fut24]. 5

AI Pitfall Methodological and conceptual errors centered around design decision-making, that, if unnoticed, result in incorrect behavior of the AI system. 5, 65

AI System Article 3 of the AI Act defines AI System as "[...] a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" [Fut24]. In addition, we interpret the AI system as a combination of all underlying processes and design decisions. Consequently, information on the intelligent system's quality can be derived from the individual lifecycle implementation, which includes linking with supplementary, contextual information. 3, 34, 357

Best Practice Proven, experience-based guidelines that enhance development processes across lifecycle stages and mitigate pitfalls. Guidelines also emerge from domain-specific needs, and are often first shared through blogs or institutional reports. Best practices remain essential for ensuring quality, safety, and efficiency in AI system design. 71

Collection-QG A type of QG within the MQG4AI building kit that defines the vertical structure of AI lifecycle stages in a flexible and modular manner. They are derived from generalizable workflows identified for executing specific lifecycle stages, resulting in a hierarchically organized QG roadmap [GM09] that bridges the gap from generic (MQG4DK) to use case-specific design (MQG4A). On higher levels, they represent a generic abstraction of the AI lifecycle, enabling the summarization and organization of associated design choices across use cases. They may comprise multiple collection- and/or leaf-QGs, which defines their QG scope. 151, 161, 316

Contributing Stakeholder Individuals actively involved in the AI system lifecycle, such as developers, domain experts, or regulators, who share responsibility of the system's impacts. We prefer the term contributing stakeholders over AI actors [Tab23, 2] to emphasize shared responsibility and ethical accountability throughout the AI lifecycle. Their role includes implementing accountable and risk-mitigated AI solutions, making design decisions, and ensuring ethical awareness. Overall, their contributions are critical to aligning systems with principles like trustworthiness and accountability. Generally, stakeholders are defined as "any individual, group, or organization that can affect, be affected by or perceive itself to be affected by a decision or activity" in ISO 22989 [ISO22a, 12]. 4, 105, 130, 159, 314

Design Decision Refers to the deliberate selection among alternative solutions to shape how a system fulfills its objectives. These decisions span from abstract architectural choices to specific implementation strategies and critically influence key system qualities such as performance, maintainability, and scalability. In the context of AI, design decisions are embedded in lifecycle design and require dynamic, context-sensitive evaluation. Competent decision-making depends on technical expertise, domain knowledge, and soft skills, among others. 71, 153, 157, 161, 173, 310, 320

Design Pattern A general, reusable solution to a recurring problem in software or ML system design. Design patterns offer structured templates that include common problems, solution strategies, trade-offs, and implementation guidance. They aim to improve maintainability, scalability, and overall quality by codifying expert knowledge into adaptable practices. [LV20]. 71

Design Science Research DSR describes the on-going and dynamic communication between design knowledge and application of that knowledge:

"[DSR] is a problem-solving paradigm that seeks to enhance human knowledge via the creation of innovative artifacts. Simply stated, DSR seeks to enhance technology and science knowledge bases via the creation of innovative artifacts that solve problems and improve the environment in which they are instantiated. The results of DSR include both the newly designed artifacts and design knowledge (DK) that provides a fuller understanding via design theories of why the artifacts enhance (or, disrupt) the relevant application contexts" [vJ20, 1]. 18, 145, 314

Domain A specific area of knowledge, activity, or expertise. In the context of science, technology, or AI, a domain is the field or subject matter to which a system or application is applied, such as healthcare, finance, or law. 16

Guiding Questions GQs are intended to support a collaborative and adaptive decision-making process that aligns with the dynamic nature of AI systems and their evolving lifecycles tailored to specific use cases. They aim to uncover essential information for RAI design decisions by encouraging stakeholders to (together) reflect on their choices and implications. We propose to base their design on the Socratic Method to foster structured knowledge elicitation, iterative discussion, and documentation, contributing to RAI KM and general AI literacy. 163, 197, 284

Information Block Component of the MQG4AI building kit designed to summarize relevant topics, essential for the implementation of RAI. It encapsulates key information derived from e.g. AI QMS requirements, as referenced in Article 17 of the AI Act [Fut24]. The MQG4AI blueprint is designed with a particular focus on system-related and risk-related aspects. Information blocks serve to highlight interdependencies across the AI lifecycle and are intended to be adaptable and modular. Where appropriate, they can be (additionally) represented as an information layer within the leaf-QG template to support a structured information flow. Overall, this modular approach is intended to facilitate integration with existing QM procedures. 125, 154, 314, 321

Leaf-QG Derived from a retrospective analysis of empirical experiments during model development, leaf-QGs are designed to process and structure design decisions through IM. These customizable information processing templates consist of modular information layers that support the integration of dynamic, lifecycle-specific insights, as well as allow for additional layers to be appended. This unifying information structure facilitates reliable, transparent, and context-sensitive documentation within MQG4DK and supports the application of use case-specific MQG4A variants. They are appended at various levels to collection-QGs, hierarchically along the

AI lifecycle, and provide means to connect vertical with additional horizontal interdependencies. Towards bridge the gap to use case-specificity, we propose to include the Socratic Method with their design, resulting in guided questions. At their core, leaf-QGs follow a three-fold structure: Content definition, Method extraction, and stakeholder-oriented Representation. An additional Evaluation layer supports the identification of open questions concerning the chosen method. Further, by highlighting relevant input and output information, leaf-QGs enable a comprehensive interdependency analysis, which is crucial for identifying and mitigating AI pitfalls by design.. 22, 151, 155, 163, 316

MQG4AI A dynamic and decentralized lifecycle planning blueprint designed to support RAI creation across the AI lifecycle. Aligned with the decentralized and continuously evolving design knowledge base (MQG4DK), MQG4AI enables the creation of use case-specific lifecycle templates (MQG4A) through a modular IM structure that allows for the integration of proven design guidelines, e.g. in the form of design pattern, or standards. Its design allows for fine-grained, transparent, and continuous information extraction and linking, where each lifecycle decision contributes to the overall quality of AI systems. Serving as a foundation for lifecycle transparency, MQG4AI aims to guide the transition from general design principles to context-specific implementation under the constant consideration of real world conditions. 3

MQG4Application MQG4A results in a combination of different lifecycle planning template versions, providing shared access to contributing stakeholders in individual, private application scenarios. They are based on MQG4DK, the living lifecycle blueprint, and follow a branching structure with one main branch and different versions for shared lifecycle exploration and concept traceability. 146, 169, 310, 319

MQG4DesignKnowledge MQG4DK is a publicly accessible living lifecycle blueprint that is intended to grow in a decentralized manner to incorporate RAI design knowledge, continuing MQG4AI, as detailed in this thesis. 146, 169, 319

Notified Body In Article 3 of the AI Act, they are defined as "[...] a conformity assessment body notified in accordance with this Regulation and other relevant Union harmonisation legislation" [Fut24]. They carry out "[...] third-party conformity assessment activities, including testing, certification and inspection", aiming to assess "[...] whether the requirements [...] relating to a high-risk AI system have been fulfilled" [Fut24]. 4, 111, 132

Provider Defined as any body "[...] that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge" in Article 3 of the AI Act [Fut24]. 4, 111, 130

QG Application In the context of MQG4AI, QG application supports quality by design through decentralized, structured, and stakeholder-specific IM. Positioned at key lifecycle transitions, QGs combine elements of standardized *quality guidelines* based on a generic AI lifecycle, and adaptable *quality strategies* tailored to project-specific needs and e.g. guided by the Socratic Method. Leaf-QGs capture design decisions using a layered template that enables interdependency linking across horizontal and vertical dimensions and incorporates guiding questions. This allows both collection-QGs and individual leaf-QGs to structure and evaluate RAI information depending on the individual interaction scenario, whether focused on method-specific contributions (MQG4DK) or applied project integration (MQG4A). 162

QG Criteria Define measurable expectations for AI lifecycle stages and design decisions, supporting risk mitigation and RAI implementation. In MQG4AI, criteria evolve with lifecycle granularity, from generalizable design guidelines at higher levels to concrete requirements at leaf-QG level. Derived from official guidelines such as standards or expert input, QG criteria guide whether design decisions meet stakeholder expectations. Leaf-QGs document these decisions along the AI lifecycle using a unifying information processing template and are flexibly extended, as illustrated for RM, to accommodate context-specific needs and compliance with AI regulation. Within MQG4AI, QG criteria evolve in alignment with MQG4DK, aiming to ensure on-going high-quality within a dynamic real world. This results in a multi-stakeholder approach by design. 163

QG Evaluation Refers to the structured analysis of information flows across QGs within the AI lifecycle and in relation to contextual information integration. Based on the QG interdependency graph, flexible and context-specific evaluations are enabled. This IM and KM-based evaluation, which highlights interdependencies and displays similarities to *roadmapping* [GM09], forms the basis to support AI QM through traceable, lifecycle-aligned decision-making. 165

QG Gate-keeping Gate-keeping refers to the decision-making process at each QG, assessing whether an AI system adheres to quality criteria in form of design guidelines and IM. Gate-keeping is decentralized and informed by the completeness and content quality of filled leaf-QG layers, as well as the presence and coherence of collection-QGs. Leaf-QGs follow a layered structure, which is envisioned to support (semi-)automated assessment, aligned with the AI Act's risk-based approach. 163

QG Inheritance Originating from object-oriented programming (OOP) [Par20], inheritance is used in MQG4AI to formalize design decisions that share core logic but diverge in complexity or specificity. For example, multi-label classification inherits from binary classification by extending the same underlying calculations (e.g. per-class probabilities) using averaging techniques. This inheritance supports design knowledge reuse and modular structuring within MQG4DK, allowing consistent adaptation across related AI system scenarios. For instance, we propose to adjust QG naming (*name; view*) within MQG4DK accordingly, resulting in a shared name and differing views. 161, 281

QG Interdependency Graph A bidirectional, hierarchical graph structure organizing collection- and leaf-QGs along the AI lifecycle, based on their roadmap [GM09], as well as identified input and output information for design decision-making. It captures vertical and horizontal interdependencies across generic and use case-specific stages, as well as links contextual system information. This enables comprehensive and dynamic information extraction, linking, as well as structured design decision management. 161

QG Naming Refers to a standardized naming convention for leaf-QGs and is broadly defined as: *design decision's name; view on its applicability*. It organizes design decisions according to their applicability. This may relate to recurring AI task structures (e.g. classification, segmentation). Overall, it should be aligned with a reasonable AI use case organization that addresses structural similarities, and how they relate to structured design decision-making. Overall, it is intended to facilitate lifecycle traceability, highlights interdependent design decisions, and organizes contributions within MQG4DK. The identification of a reasonable approach to QG naming within MQG4A is part of future work. 148, 202, 248, 281

QG Positioning Defines the hierarchical placement of collection- and leaf-QGs along the AI lifecycle, which could be called roadmapping [GM09]. Collection-QGs result in generalizable stage structures with varying degrees of generality (e.g. Data, Configuration, Evaluation), while accordingly appended leaf-QGs document concrete design decisions. Overall, they support design, planning and documentation of RAI system creation. 161

QG Relations These relationships capture how QGs build upon, extend, or integrate with one another. They describe structured connections between QGs within MQG4AI, supporting modular knowledge organization in MQG4DK. Possible relations, inspired by object-oriented principles (e.g. inheritance, aggregation or composition) [Par20], may further define how QGs interact. Identifying and formalizing these relations is part of ongoing and future work to enhance design knowledge traceability and reuse. Understanding how QG Relations influence the generation and configuration of MQG4A template versions is equally a subject of future research. 161, 281, 322

QG Scope Depends on the vertical and horizontal interdependencies of a QG within the AI lifecycle. Vertical scope reflects lifecycle stage hierarchy, and higher-level collection-QGs are defined by aggregating the horizontal interdependencies of lower levels, which are derived from the information structures of leaf-QGs. They reflect horizontal scope, which captures dependencies on contextual resources and output/input relationships with other QGs. 161

QG Scoring System A multi-index evaluation framework envisioned to assess AI system quality based on information structured around QGs. It supports tailored evaluation scenarios, deriving metrics like a *Compliance Index*, which may aggregate multiple evaluation outcomes into a single score, or a *Multi-Stakeholder Fairness Score*, based on the inclusion and representation of stakeholders across QGs. The scoring system promotes transparency by linking each index to lifecycle data, documented in the MQG4AI blueprint. It can be used for comparative analyses across use cases, contributing to RAI creation and governance. 167, 321

QG Tags Structured metadata information layers assigned to leaf-QGs to enable intelligent, use case-specific retrieval of design guidelines from MQG4DK, resulting in MQG4A-v0. We propose to define them in accordance with ML design pattern characterization [WH22], resulting in the following structure: *Name, Intent, Problem, Solution, Applicability, Consequences, Usage Example*. It aims to support the targeted configuration of MQG4A through intelligent tag-based queries. 148, 169, 203, 248, 282

Quality Gate A central building block within MQG4AI. Quality Gates serve as a modular checkpoint for managing information across the AI lifecycle. Therefore, they reflect project-specific AI lifecycles similarly to a "digital twin" (MQG4A) based on a modular and hierarchical information structure. Their high-level organization with varying degrees of more use case-specific design guidelines is derived from a living RAI lifecycle blueprint (MQG4DK). Two primary types exist: collection-QGs, which define and structure the AI lifecycle, and leaf-QGs, which document concrete design decisions that require more complex decision processes. 21, 151, 159, 314

RAI Design Knowledge Comprises all existing and novel design knowledge required to implement and evaluate intelligent systems in a manner that is lawful, ethical, and accountable. This includes both technical and contextual information, with the aim of enabling successful integration into real world settings. It comprises implicit, explicit, and tacit knowledge. Throughout this thesis, the terms *RAI/AI knowledge* may also refer to this concept, though it should not be confused with the knowledge an AI system has learned or inferred, as described in ISO 22989 [ISO22a, 18]. 5, 112, 114, 119, 159

RAI Information Management Describes MQG4AI's functionality and refers to the structured process of collecting, organizing, maintaining, and disseminating RAI design knowledge across the AI system lifecycle. Within the MQG4AI blueprint, this is achieved through generic and customizable building blocks, supporting transparency, traceability, and informed decision-making for lawful, ethical, and accountable AI system design. In alignment with ISO 5338, it ensures that implementation-relevant information, such as technical details, project context, stakeholder responsibilities, and risk considerations, is reliably managed and made accessible to designated contributing stakeholders [ISO23c]. 160

RAI Knowledge Management Refers to the systematic organization, transformation, and application of knowledge related to the implementation of RAI. MQG4AI is intended to incorporate RAI KM through IM. This enables the continuous generation and application of RAI knowledge by converting abstract, often human-dependent design knowledge into tangible and measurable formats through structured information processing. Within MQG4AI, IM acts as the mechanism through which knowledge (or possibly, scattered information) is collected, externalized and made actionable, allowing it to be reused, validated, and extended across the AI system lifecycle. [ISO23c] Together, these processes support AI literacy, stakeholder collaboration, and sustainable, high-quality AI system design through iterative knowledge refinement and contextual adaptation. 115

Risk Probability of occurrence of harm and the severity of that harm combined. Harm focuses on human beings and the planet, in contrast to harm to the organization. [Eura]. 11, 91

Risk Analysis Comprises the identification and estimation of risks regarding their severity and likelihood. This entails identifying risk sources and related events that may result in harm based on the intended use and reasonably foreseeable misuse, resulting from human behavior or interaction with other systems [Fut24]. 91

Risk Control Process for reducing or maintaining risk at acceptable levels through specific measures and decisions. These measures can target various aspects of the risk analysis process, such as risk sources or enabling events, by reducing their likelihood, severity, or both. 95

Risk Evaluation Identified and estimated risks are evaluated, and the RMS defines criteria for risk acceptability assessment. Risk evaluation is an iterative process that assesses both residual and newly introduced risks after implementing control measures, repeating until the residual risk is deemed acceptable. If residual risk is considered unacceptable and further mitigation is not feasible, a benefit-risk analysis [ISO19, 14] can aid in decision-making by comparing the benefits of the intended use with the remaining risks. 94

Sector In contrast to *domain*, which often relates to knowledge or subject areas, *sector* emphasizes organizational or economic divisions and is often used in policy, business, or industry contexts. 16

Socratic Method Question-driven technique that fosters critical thinking by helping individuals examine their assumptions and reasoning through guided dialogue. Traditionally used in education, it shifts the responsibility of learning to the participant and promotes reflective, self-directed inquiry. [IA12] In the context of MQG4AI, this method inspires the use of Guiding Questions (GQs) within leaf-QGs. 197, 284

Standard A formally established set of guidelines that ensures consistency, quality, and interoperability across systems. Standards provide structured frameworks that support best practices and explicitly address the interface between technical implementation and regulatory compliance. 72

Acronyms

3D Three-dimensional. xv, 30–32, 35, 221, 224, 226–232, 234, 235, 286, 289–291, 293

Adam Adaptive Moment Estimation. 44, 265, 303

AI Artificial Intelligence. v, vi, xi–xv, 2–22, 24–30, 32–41, 45–56, 59, 62, 64–72, 74–91, 93–142, 144–147, 149–173, 175–178, 181–212, 216–219, 222, 231, 232, 234, 235, 237–239, 242, 243, 246, 248, 249, 252–254, 256, 257, 259–266, 268–270, 276, 280, 283–287, 289, 304, 306, 310, 312, 314–324, 357–364, 372, 374–379

AIRO AI Risk Ontology. 91–93, 95, 119, 121, 125, 127, 129, 130, 132, 262, 317

AL Active Learning. 25, 29, 218

ALTAI Assessment List for Trustworthy AI. 13, 98, 127, 134, 253, 257, 290, 318

BM Benefit Matrix. 277, 278, 283

CAM Class Activation Mapping. 56, 218

CDSS Clinical Decision Support. 31, 34, 122, 142, 185, 212, 231, 235, 240, 262

CE Cross Entropy Loss. 44, 265, 299

CEN European Committee for Standardization. 82, 85–87

CENELEC European Committee for Electrotechnical Standardization. 82, 85, 86

CIST Category Imbalance and Cost-Sensitive Thresholding. 266, 274, 276, 278–280

CNN Convolutional Neural Network. 35, 42, 218, 265

CT Computed Tomography. 42, 66, 117, 209, 215, 231

CVD Cardiovascular Diseases. 212–214

DICOM Digital Imaging and Communications in Medicine. 38, 142

DL Deep Learning. 32, 34–36, 39–45, 65, 67, 72, 124, 130, 205, 218, 266, 298, 322, 324

DNN Deep Neural Network. 2, 3, 5, 6, 9, 10, 19–21, 29, 33, 35, 36, 39–45, 51, 53, 55, 56, 59, 60, 65, 66, 74, 97, 113, 114, 124, 139, 178, 185, 218, 241, 315

DSR Design Science Research. v–viii, 18–21, 114, 118, 120, 124, 135, 142, 144, 145, 150, 154, 157, 158, 198, 202, 238, 311, 315, 319, 322, 358, 359

ECG ElectroCardioGram. xiv, 21, 22, 26, 29, 30, 47, 49, 50, 55, 56, 68, 124, 157, 177, 183, 204, 205, 211, 212, 214–221, 235, 256–258, 260, 262, 264, 266, 267, 270, 271, 274, 276, 277, 280–283, 319

EHR Electronic Health Record. 38, 210

EM Emergency Medicine. 173, 183, 204, 205, 212, 217, 219, 221, 256–258, 262, 264, 267, 277, 281–283, 319

EndoFLIP Endolumenal Functional Lumen Imaging Probe. 224, 227

EU European Union. v, 3, 6, 10, 12, 13, 20, 38, 39, 65, 75–77, 82, 85, 86, 91, 97, 98, 106, 111, 119, 120, 127–129, 132, 134, 147, 151, 175, 210, 214

FHIR Fast Healthcare Interoperability Resources. 38, 209

FMTI Foundation Model Transparency Index. 49, 50

FPR False Positive Rate. 271, 275, 279

GDPR General Data Protection Regulation. 38, 75, 177, 205, 210

GenAI Generative AI. 30, 31, 35, 36, 47, 69, 71, 131, 150, 158, 234, 235, 324

GPU Graphics Processing Unit. 201, 234, 291

GQ Guided Questions. 197, 198, 248, 359, 365

GUI Graphical User Interface. 30, 187, 207, 243

HARHD Heart Arrhythmias/Rhythms and other Heart Diseases. 215–220, 260, 263, 266, 267, 274, 277

HLEG High-Level Expert Group appointed by the European Commission. 3, 51, 77, 100–104

HRM High-Resolution Manometry. 224, 225, 227, 229–231

IEC International Electrotechnical Commission. 79, 80, 82–90, 97–99, 111, 120, 123, 167, 169, 183, 201, 257, 259, 260, 269, 272, 321

IEEE Institute of Electrical & Electronics Engineers. 79, 82, 84, 175, 186, 237, 253, 254, 319

IG-NB German Notified Bodies Alliance. 71, 79

IM Information Management. v–viii, 2, 5, 12, 13, 72, 75, 99, 110–115, 118, 119, 125, 130, 134, 141, 142, 144, 146, 147, 150, 153, 157, 159–164, 167, 170, 172, 201, 255, 281, 286, 314, 315, 317, 321, 359–361, 364

IoU Intersection over Union. 296, 300, 304

ISO International Organization for Standardization. 79, 80, 82–99, 111, 113, 115, 120–123, 125, 126, 130–132, 167, 169–172, 175, 181, 183, 188, 191, 194, 195, 201, 202, 205, 206, 257, 259, 260, 268, 269, 272, 284, 285, 317, 321, 358, 364

IT Information Technology. 31, 84

IVDR In Vitro Diagnostic Medical Device Regulation. 16, 75–77, 89, 205, 210

KM Knowledge Management. 10, 110–113, 115–118, 125, 133, 135, 146, 153, 164, 170, 198, 286, 288, 359, 361, 364

LES Lower Esophageal Sphincter. 221, 222, 224, 229

LIME Local Interpretable Model-agnostic Explanations. xv, 47, 52, 56, 57, 60, 63, 156, 186, 187, 197, 237, 240, 242, 244, 246–250, 252, 319

LLM Large Language Model. 31, 34, 35, 89, 132, 198

LSTM Long Short-Term Memory. 42, 43, 218, 265

MCC Matthew's Correlation Coefficient. 271, 279

MDR Medical Device Regulation. 16, 17, 75–77, 89, 91, 94, 95, 205, 210, 211

ML Machine Learning. 5, 35, 37, 38, 41, 44, 48, 50, 51, 55, 59, 62, 68, 72, 86, 120, 192, 203, 207, 209, 244, 250, 259, 261, 264, 282, 312, 322, 363, 372

MLOps Machine Learning Operations. 170, 172, 173, 192

MQG4A MQG4Application. vi, viii, xiii, xiv, 21, 22, 25, 146–151, 153, 155–157, 162, 163, 167–169, 171–174, 180–186, 189, 190, 196, 199–203, 211, 234, 235, 237, 242, 247, 248, 250, 251, 253, 254, 256, 262, 266, 269, 270, 282, 284–289, 291, 297, 298, 302–304, 306–312, 314, 315, 318–323, 358–363

MQG4AI Generic and Customizable Methodology based on Quality Gates towards Certifiable AI in Medicine. v–viii, xii, xiii, 3, 18–22, 24, 25, 32, 33, 47, 72, 74, 75, 90, 96, 99, 105, 109–129, 133–137, 139–165, 167–171, 173, 174, 177, 179, 181, 184, 186, 189, 195–198, 201–205, 211, 217, 235, 237, 238, 246, 247, 253–257, 264, 269, 280, 281, 286, 298, 310–312, 314, 315, 317–324, 358–365, 372, 373, 375, 377

MQG4DK MQG4DesignKnowledge. vi, viii, xiii, xiv, 21, 22, 25, 146–151, 153, 155–157, 161–163, 167, 169, 171, 173, 180, 182–187, 189, 190, 196–203, 211, 217, 219, 235, 237, 238, 242, 243, 246–248, 250, 253–257, 262–264, 266, 270–272, 276, 280–285, 287–289, 308, 310–312, 314, 315, 319–323, 358–363

NDCG Normalized Discounted Cumulative Gain. 251, 253

NIST American National Institute of Standards and Technology. 13, 82, 97–101, 103–105, 108, 122, 135

NLP Natural Language Processing. 43, 84, 205

OECD Organization for Economic Co-operation and Development. 97–100, 102, 105, 120, 135, 169

OOP Object-Oriented Programming. 161, 281, 322

POEM PerOral Endoscopic Myotomy. 223–225

PPV Positive Predictive Value. 258, 259

PR AUC Precision-Recall Area under the Curve. 259, 260, 266, 271, 272, 278, 280

QG Quality Gate. v, vii, viii, xiii–xvii, 21, 22, 110, 114, 120, 125, 133–135, 137–139, 141, 146–149, 151–169, 171–193, 195–204, 211, 219, 235, 237, 238, 242–258, 260, 264, 265, 267, 269, 270, 276, 280–289, 294–300, 302–304, 306–312, 315–323, 358–363, 365

QM Quality Management. v, vii, 5, 7, 9, 15, 17, 20–22, 74, 75, 79, 82, 83, 85, 87–90, 99, 119, 120, 125, 141, 151, 153, 157–159, 162, 164, 165, 170, 171, 193, 195, 199, 315, 317, 318, 323, 359, 361

QMS Quality Management System. xiii, 3, 4, 6, 8, 75, 76, 79, 85, 87, 89, 90, 95, 99, 110, 111, 147, 150–152, 164, 201, 210, 321, 359

RAI Responsible AI. v–viii, xii, xiii, 2, 4–6, 8–15, 17–24, 26, 28, 30, 33, 34, 36, 46, 49, 51, 53, 65, 68, 70–72, 74, 75, 77–79, 87, 90, 99, 101, 104–121, 124, 125, 127–130, 133–137, 139–142, 144–150, 153, 154, 156–161, 163, 167, 169, 173, 183, 187, 189, 190, 195, 198, 199, 201–203, 205, 206, 208, 210, 238, 247, 248, 250, 253, 255, 257, 260, 263, 269, 286, 290, 311, 312, 314–319, 321–323, 359–364, 372, 373, 375–377

RCA-IUnet Residual Cross-spatial Attention guided Inception U-Net. xvi, 298, 302–305, 311, 312, 320

ReLU Rectifier Linear Unit. 35, 42

RM Risk Management. 7, 12, 13, 15, 18, 21, 46, 51, 66, 74, 75, 82, 83, 87–91, 94, 95, 97–100, 104, 105, 108–111, 119, 120, 123, 125, 127, 130, 134, 135, 137, 141–143, 149, 151, 152, 154, 158, 161, 164, 166, 167, 170, 181, 191, 193, 201, 210, 256, 257, 283, 284, 288, 289, 294, 295, 297, 307–309, 314, 317, 318, 321, 361

RMF Risk Management Framework. 98–100, 105, 135

RMS Risk Management System. 4, 46, 76, 90, 94, 96, 101, 104, 111, 128, 137, 144, 151, 364

RNN Recurrent Neural Network. 42, 43, 218

ROC AUC Receiver Operating Characteristic Area under the Curve. 51, 258–260, 264, 266, 271, 272, 274, 278, 280

SGD Stochastic Gradient Descent. 43, 44, 299

SHAP SHapley Additive exPlanations. xv, 47, 52, 57, 58, 60, 156, 186, 187, 197, 237, 240, 242, 244, 246–250, 252, 319

SNOMED CT Systematized NOmenclature of MEDicine – Clinical Terms. 221, 266, 268

STEM Science, Technology, Engineering and Mathematics. 29, 376

TAI Trustworthy AI. v, viii, 3, 4, 7, 11–13, 49, 51, 54, 77, 78, 82, 83, 93, 98–106, 108, 109, 119, 122, 130, 134, 135, 141, 142, 144, 153, 154, 166, 167, 170, 201, 253, 256, 283, 290, 318, 321, 375

TBE Timed Barium Esophagogram. xvi, 31, 211, 224, 228–231, 234, 288, 289, 291, 293–295, 297, 298, 302–304, 306–308

TE Tubular Esophagus. 222, 224

US United States. 82, 132, 212, 266

WHO World Health Organization. 213

XAI Explainable AI. xii, 5, 12, 33, 46, 47, 51–61, 63–65, 68–70, 84, 139, 156, 185, 186, 202, 218, 237, 238, 240, 241, 243–245, 248, 254, 255, 273, 284, 319, 375

A
Reflecting on Mentorship

Generally, I am convinced that seeking out and providing mentorship immensely benefits one's personal and professional evolution. Throughout my PhD journey, I have the opportunity to connect with three incredibly inspiring mentors who are guiding me along the way, significantly impacting the creation of MQG4AI. This section honors our collaboration.

Prof. Dr. Sophie Weerts, at the Faculty of Law, Criminal Justice and Public Administration,[1] has been a major source of inspiration for the creation of this thesis, for which I am eternally grateful. Our time together at her lab, visiting the Institut de Hautes Études en Administration Publique (IDHEAP) in Lausanne, was both enriching and inspiring. During my visit in April 2024, I had the privilege of presenting an early version of the concept to an audience from a law background. Sophie introduced me to *Design Science Research* [vJ20], as well as the foundational definition of RAI in the context of this thesis [DRN23]. Her perspective also influenced our previously introduced publication *Responsible AI, ethics, and the AI lifecycle: how to consider the human influence?* [EM25b]. She continues to offer her invaluable professional legal perspective on the regulatory landscape surrounding AI in Europe and I am looking forward to sharing perspectives and insights along the way, including guidance for future steps. And, thankfully, she encouraged me to continue seeking mentorship, which now has became an integral part of my journey.

Dr. Jauwairia Nasir, a postdoctoral researcher who focuses on socially assistive robots, multi-modal behavioral analytics, data-driven modeling, and applied ML in education and healthcare,[2] has been a precious source of inspiration and an invaluable guide along the way. We met over the Women in AI global network website[3], that I found online, looking for mentorship. Since we both work at the University of Augsburg, we connected, and quickly became close friends and trusted collaborators. In addition to wonderful conversations, fruitful introductions, walks in nature, and invitations to scientific events, Jauwairia always

[1] `https://applicationspub.unil.ch/interpub/noauth/php/Un/UnPers.php?PerNum=1198672`

[2] `https://jauwairianasir.com/`

[3] `https://www.womeninai.co/wailabs`

shared her PhD experience. I am excited to continue this journey together, and who knows, maybe we will even organize a podcast or launch new projects together, connecting our professional insights.

Emma Grönvik Möller, LL.M., the CEO and founder of *Lumiera*,[4] recently became an invaluable mentor, offering her deep perspective on the intersection of RAI and the economy. Her insights have profoundly shaped my thinking. Emma has encouraged me to evaluate the practical applicability of the concepts presented in the present work, helping guide the methodology's evolution. Finally, I am incredibly grateful to Emma for providing me with guidance on how to approach the next step after completing my PhD, and, following her invitation to Lisbon, I am exploring a multitude of options, which include on-going considerations surrounding MQG4AI. The idea to refer to MQG4AI as *blueprint* was equally inspired during this visit, and I am incredibly grateful for our engaging and thoughtful conversations. Looking forward to more experiences and learning together.

Female High Potentials Mentoring (FemHighPot)

From autumn 2023 to autumn 2025, I had the honor of participating in the FemHighPot mentorship program.[5] The program supports women on academic career paths and encourages us to build strong professional networks, including actively seeking out mentorship. Our cohort, which consists of 12 PhD candidates from a variety of disciplines at the University of Augsburg, offers a fantastic foundation for collaboration and exchange. Sharing interdisciplinary perspectives is essential for research, and contributed to the creation of MQG4AI. In addition to kick-starting our personal mentoring journey, FemHighPot offers a series of workshops on key topics such as core competencies for academic careers, visibility and career planning, press relations, and science communication. Overall, team-building activities and shared experiences have laid the foundation for a supportive and empowering network where we continue to grow together.

[4] `https://www.lumiera.ai/`

[5] `https://www.uni-augsburg.de/de/verantwortung/gender-equity-diversity/gender-equity/service/mentoring-female-high-potentials/`

B

Reflecting on Science Communication

Science communication is a vital part of being a scientist. Engaging in various events with diverse audiences and goals has significantly shaped my professional journey and the creation of this thesis. I actively participate in public dialogue to raise awareness and facilitate discussions around topics concerning (medical) AI, helping to bridge technological developments and societal understanding. The aim is to make complex ideas accessible and relevant. The following selection presents a broad collection of diverse event formats, before highlighting STEM education.

- *Interdisciplinary Panel Discussion* In June 2023, I had the honor of joining a panel discussion on Interdisciplinarity and Health Humanities during the inaugural event of the Graduate Center at the Faculty of Law at the University of Augsburg. The event began with a powerful keynote by Prof. Dr. Kirsten Ostherr on *Translational Medical Humanities: Connective and Transdisciplinary Scholarship for Global Challenges* [Ost23]. Together with colleagues from the humanities, medical, and social sciences, we explored the evolving role of healthcare and the societal impact of AI. The conversations highlighted the importance of interdisciplinary collaboration and the ethical dimensions of technology in medicine, directed at a public audience.

- *Long Night of Science* At the Long Night of Science 2024,[1] we presented our student research project on the digital support of clinical research, diagnosis, and treatment of the rare disease Achalasia [EM23] at the CReAITech booth. Overall, the evening was a great success and provided plenty of opportunities for exchange with a diverse audience. Together with participating students, we had a multitude of discussions with the broad public reaching from ethical and AI-related questions to medical ones.

- *Königsbrunner Campus* As further contribution to our ongoing project on the technical enhancement of research, diagnosis, and treatment of the rare dis-

[1] https://www.youtube.com/watch?v=FZ9gtCeeHjY&ab_channel=Universit%C3%A4tAugsburg

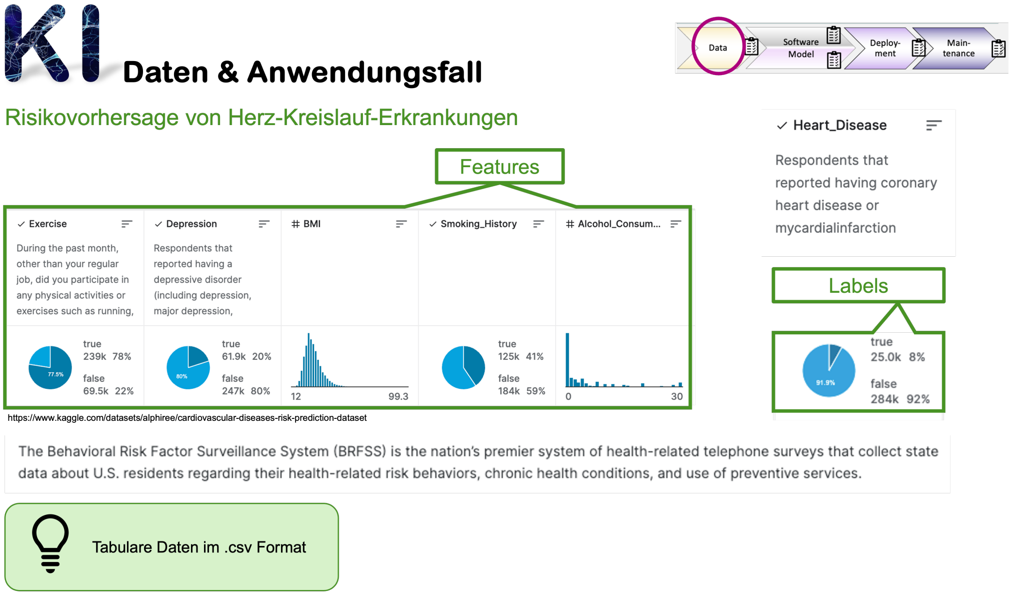
ease Achalasia [EM23], in 2024, our team presented key developments, that are introduced in section 5.2.2. In collaboration with the University Hospital Augsburg and the CReAITech, we showcased our interdisciplinary approach combining embedded ethics, AI-driven design of our prototypical software, and long-term implementation strategies. The audience consisted mostly of Achalasie patients and related friends and family, which contributed a unique perspective to the follow-up discussion and emphasized the need for advancing research on Achalasia.

- *Scientific Exchange at IDHEAP* As part of a research visit to the Institut de hautes études en administration publique (IDHEAP) in Lausanne in 2024, an interdisciplinary exchange took place with a focus on certifiable AI. The session brought together primarily legal scholars, offering a more technical and scientifically complex science communication setting compared to previous public-facing presentations. The discussion centered on an earlier version of this thesis. The presentation aimed to make technical content accessible to a legal audience while fostering dialogue across disciplinary boundaries and highlighting the importance of tailoring science communication to expert, yet non-technical, stakeholders in the broader regulatory discourse on RAI, which is lawful, ethical, and accountable AI [DRN23].

- *Guest Lecture* As part of the seminar kickoff on *Explainable AI & Ethics* at the University of Applied Sciences Kempten, I presented an early version of this thesis, focusing on TAI, following an invitation by Prof. Dr. Rafael Mayoral. Drawing on examples from the medical domain, the presentation highlighted both the specific potentials and challenges of implementing AI in healthcare. A central focus was placed on the role of XAI across the AI lifecycle, illustrating how transparency and interpretability impact and contribute to the development of responsible, certifiable technologies. The lecture was tailored to university students in higher education, particularly those with a background in computer science and related disciplines. A critical discussion on how to design AI systems that meet the standards of trustworthiness in high-stakes applications rounded up the event.

STEM Education plays a crucial role in fostering digital literacy and critical thinking, especially within the context of AI. As intelligent systems continue to shape various industries, the need for informed and intelligent users becomes even more critical, complementing MQG4AI's vision.

We have developed educational content and organized events that aim to strengthen foundational knowledge in AI, ensuring that learners are equipped to engage responsibly and meaningfully with intelligent tools. We aim to lay the foundation for an infrastructure that facilitates the communication of knowledge and connects scientific personnel and materials with schools, promoting collaboration between academia and education. Events are related to the Faculty of Ap-



KI Daten & Anwendungsfall

Risikovorhersage von Herz-Kreislauf-Erkrankungen

✓ Exercise: During the past month, other than your regular job, did you participate in any physical activities or exercises such as running,

✓ Depression: Respondents that reported having a depressive disorder (including depression, major depression,

✓ Heart_Disease: Respondents that reported having coronary heart disease or mycardialinfarction

https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services.

Tabulare Daten im .csv Format

Figure B.1.: An example of our approach to resource creation for STEM education. Here, we introduce basic machine learning concepts to interested students during a summer holiday course on AI.

plied Computer Science[2] and STEM Education[3] at the University of Augsburg.

For instance, we developed engaging materials to support knowledge transfer at the interface of AI and medicine. These include short Reels addressing AI in medicine from a PhD-perspective in the series *AI_ME*[4] for accessible science communication on social media. Further, we began a university-internal resource collection for STEM education. Materials comprise introductory content that can be easily presented by different specialists. For instance, we held a workshop on machine learning in medicine, designed to introduce students (and university audiences) to core concepts, applications, and ethical considerations of AI in healthcare, as illustrated in Figure B.1. In various outreach events, we either used the developed materials or created new, tailored slide decks to engage different audiences. These events included the *Career Day* for senior grades at the Rudolf-Diesel Gymnasium in Augsburg, where we introduced AI and society as a potential career path, further, we provide *Einblicke in Forschung und Lehre* (insights into science and teaching) at an event organized by the Faculty of Computer Science at the University of Augsburg, showcasing research and teaching practices, and presenting our research project on Achalasia, the rare disease to students. Another highly appreciated event is the *Girls' Day*, where we introduce RAI to inspire young women in tech.

[2] https://www.uni-augsburg.de/de/fakultaet/fai/informatik/vor-dem-studium/tag-der-informatik/tag-der-informatik-2024/#:~:text=Termin%20und%20Ort%202024,dem%20Campus%20der%20Universit%C3%A4t%20statt.

[3] https://www.uni-augsburg.de/de/forschung/einrichtungen/institute/amu/bildung/

[4] https://www.instagram.com/reel/CzBDxQPriCT/?utm_source=ig_web_copy_link

C

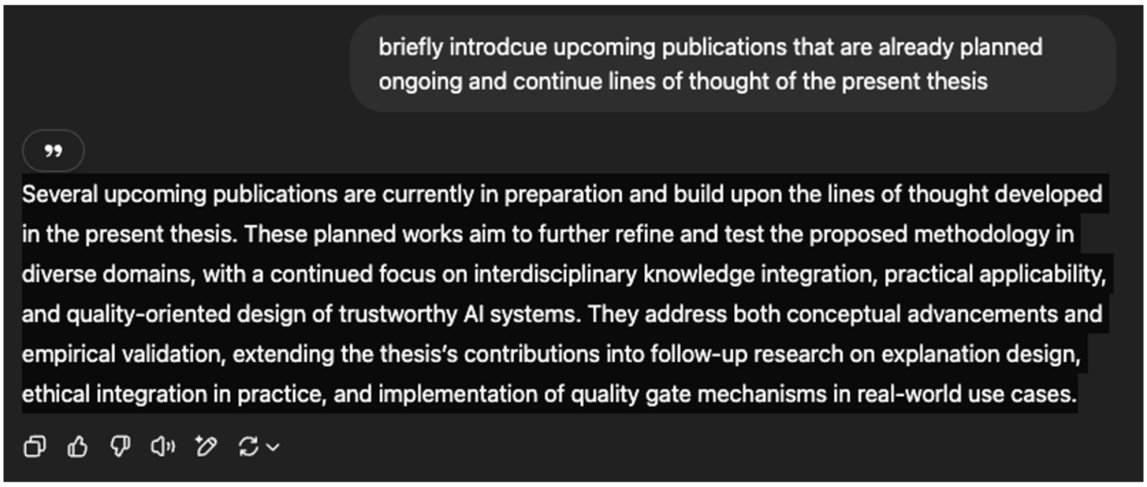
Reflecting on the Intelligently Enhanced Writing Process

In the spirit of transparency and responsible scientific participation, this section provides an open reflection on my personal experience with AI tools during the creation of this thesis. TAs AI systems are increasingly becoming part of academic and professional workflows, it is important to critically assess and document their use. Sharing how and to what extent such tools were involved not only supports academic integrity but also contributes to the evolving discourse on human–AI collaboration in research.

The present thesis was written between November 2024 and May 2025, with continuous progress made over time. AI tools helped me maintain a consistent writing pace, allowing me to make progress even during phases when my creative energy was lower, as illustrated in Figure C.1. In support of the writing process, Chat-GPT (public version) was used selectively as the main tool to improve readability, writing speed, and reusability of content. It was launched in 2022 and is the first tool I started experimenting with. For instance, as depicted in Figure C.2, I found it especially helpful for adapting and rephrasing notes from presentations, earlier publications, and materials within the GitHub repository[1], which was created prior to the writing phase without AI support. This repository, developed in close cooperation with explainability and RAI experts [EM25a], was foundational to this work.

It is important to note that the idea, overall storyline, structure, and content definition were developed independently of AI, in close collaboration with my supervisor. The role of AI remained strictly supportive, never single-handedly generative in terms of ideas or arguments. Every output was double checked and corrected, in alignment with MQG4AI, including outputs that Chat-GPT derived from remembering certain information of previous prompts when growing a memory. However, I also had philosophical and in-depth conversations with AI, which occasionally sparked unexpected insights and contributed to the development of intriguing ideas, such as the translation of *Transversality* to RAI [EM25b].

[1] https://github.com/miriamelia/MQG4AI/blob/main/README.md



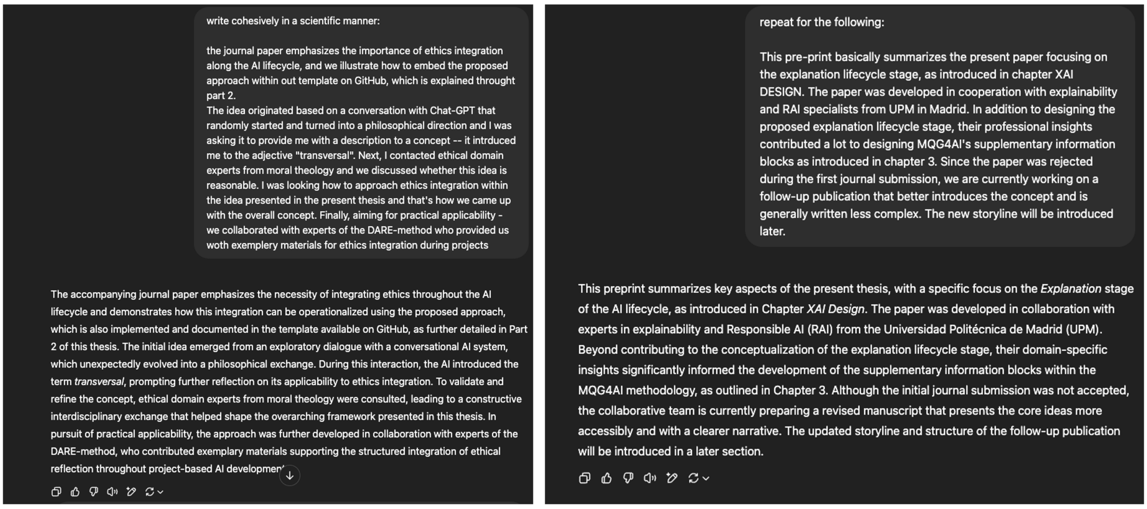
briefly introdcue upcoming publications that are already planned ongoing and continue lines of thought of the present thesis

Several upcoming publications are currently in preparation and build upon the lines of thought developed in the present thesis. These planned works aim to further refine and test the proposed methodology in diverse domains, with a continued focus on interdisciplinary knowledge integration, practical applicability, and quality-oriented design of trustworthy AI systems. They address both conceptual advancements and empirical validation, extending the thesis's contributions into follow-up research on explanation design, ethical integration in practice, and implementation of quality gate mechanisms in real-world use cases.

Figure C.1.: An example when my creative writing energy was lower and Chat-GPT supported me to produce beautiful writing. The result on page 38: "Two upcoming publications are currently in preparation and build upon the lines of thought developed in this thesis" – This example illustrates very well how the intelligent system is equipped to produce well-written text that user can choose from. This at times is easier and quicker than producing new text from scratch. Also, the output mentions *Quality Gate* independently of the direct prompt, illustrating the AI's personalization over time through growing a memory.

All passages generated with the help of AI were carefully reviewed, edited, and refined multiple times to ensure precision and alignment with the academic tone and objectives of the theses. Especially, when using AI to rephrase quotes from research papers, I made sure to translate the correct logic, which, depending on the paper's complexity, was not always trivial.

Finally, literature research was conducted predominantly without AI support. Only toward the end did I briefly explore tools like Perplexity or Consensus to complement my research strategy and for experimentation. Additionally, I experimented to a limited extent with the public versions of Claude and Gemini. All AI tools used during the writing process were publicly available versions and no enterprise or private instances were accessed. Also, I have paid attention not to reveal information that is not publicly available on the internet, and in my data control settings of Chat-GPT, the functionality to "improve the model for everyone" is turned off.



write cohesively in a scientific manner:

the journal paper emphasizes the importance of ethics integration along the AI lifecycle, and we illustrate how to embed the proposed approach within out template on GitHub, which is explained throught part 2.
The idea originated based on a conversation with Chat-GPT that randomly started and turned into a philosophical direction and I was asking it to provide me with a description to a concept -- it intrduced me to the adjective "transversal". Next, I contacted ethical domain experts from moral theology and we discussed whether this idea is reasonable. I was looking how to approach ethics integration within the idea presented in the present thesis and that's how we came up with the overall concept. Finally, aiming for practical applicability - we collaborated with experts of the DARE-method who provided us woth exemplery materials for ethics integration during projects

The accompanying journal paper emphasizes the necessity of integrating ethics throughout the AI lifecycle and demonstrates how this integration can be operationalized using the proposed approach, which is also implemented and documented in the template available on GitHub, as further detailed in Part 2 of this thesis. The initial idea emerged from an exploratory dialogue with a conversational AI system, which unexpectedly evolved into a philosophical exchange. During this interaction, the AI introduced the term *transversal*, prompting further reflection on its applicability to ethics integration. To validate and refine the concept, ethical domain experts from moral theology were consulted, leading to a constructive interdisciplinary exchange that helped shape the overarching framework presented in this thesis. In pursuit of practical applicability, the approach was further developed in collaboration with experts of the DARE-method, who contributed exemplary materials supporting the structured integration of ethical reflection throughout project-based AI developmen

repeat for the following:

This pre-print basically summarizes the present paper focusing on the explanation lifecycle stage, as introduced in chapter XAI DESIGN. The paper was developed in cooperation with explainability and RAI specialists from UPM in Madrid. In addition to designing the proposed explanation lifecycle stage, their professional insights contributed a lot to designing MQG4AI's supplementary information blocks as introduced in chapter 3. Since the paper was rejected during the first journal submission, we are currently working on a follow-up publication that better introduces the concept and is generally written less complex. The new storyline will be introduced later.

This preprint summarizes key aspects of the present thesis, with a specific focus on the *Explanation* stage of the AI lifecycle, as introduced in Chapter *XAI Design*. The paper was developed in collaboration with experts in explainability and Responsible AI (RAI) from the Universidad Politécnica de Madrid (UPM). Beyond contributing to the conceptualization of the explanation lifecycle stage, their domain-specific insights significantly informed the development of the supplementary information blocks within the MQG4AI methodology, as outlined in Chapter 3. Although the initial journal submission was not accepted, the collaborative team is currently preparing a revised manuscript that presents the core ideas more accessibly and with a clearer narrative. The updated storyline and structure of the follow-up publication will be introduced in a later section.

Figure C.2.: An example how AI helped me with professional writing, illustrated for the previously introduced publications related to this thesis. Overall, the text is an example for a human-machine collaboration in scientific writing.