

# Leiden University

## ICT in Business and the Public Sector

Designing an Auditing Framework for Audits of AI-Driven Risk Assessment Models used in Health and Life Insurance under the AI Act


| | |
|---|---|
| Name: | Jort Visser |
| Student ID: | s2927993 |


| | |
|---|---|
| Date: | 31/07/2026 |

| | |
|---|---|
| 1st supervisor: | Prof. dr. Mirjam van Reisen |
| 2nd supervisor: | Prof.dr.ir. Joost Visser |
| Company supervisor: | Vangjel Gjorgjiev |

MASTER'S THESIS


Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

**Abstract**

**Background:** The Artificial Intelligence Act (AIA) and growing European investments in Artificial Intelligence (AI) have increased the need for AI assessments by third-party audit firms. However, regulatory guidance remains limited, and there is a lack of shared proper AI audit methodologies for auditing under the AIA. These methodologies should be based on the Responsible AI principles and should use measurable audit criteria. AI-driven risk assessment models are often used for pricing in health and life insurance and require proper assessments regarding transparency, privacy, and fairness.

**Aim:** This study investigates how an AI audit framework used by a third-party audit firm can be improved to effectively audit AI-driven risk assessment models used in health and life insurance under the AIA.

**Method:** We carried out this research during an internship at KPMG. We conducted a literature review and two rounds of expert interviews. We used the method of Design Science Research to design the AI audit framework. The requirements and specifications for the AI audit framework were derived from the literature review and the exploratory interviews. After the framework was designed, we evaluated it through a mock demonstrator case. During these evaluation interviews, the experts assessed the framework's relevance, usability, completeness, understandability, logic, and added value.

**Results:** We developed the AI audit framework by implementing controls regarding the high-risk obligations of the AIA, relevant to providers of high-risk AI systems. Focusing on AI-driven risk assessment models used in health and life insurance, we found that fairness, bias, and data drift are crucial aspects to audit. The framework scored well above sufficient on the evaluation criteria, although several opportunities for improvement were identified. It is important to correctly determine the scope of the audit, implement measurable controls, and ensure IT auditors receive proper AI-related training.

**Conclusion:** The upgraded AI audit framework can be considered as a solid starting point for conducting AI audits under the AIA. For other AI systems besides risk assessment models used in health and life insurance, the framework needs to be tailored to the specific scope. Finally, the framework will require refinement and validation after additional guidance regarding AIA compliance becomes available from the EU.

# Contents


Abstract . . . 1

**1 Introduction** **3**
1.1 Problem statement . . . 3
1.2 Research questions . . . 4
1.2.1 Sub-questions . . . 4
1.3 Overview of the thesis . . . 5

**2 Background and related work** **6**
2.1 Artificial intelligence . . . 6
2.1.1 Black box model . . . 6
2.1.2 AI-driven risk assessment model . . . 6
2.1.3 Responsible AI . . . 7
2.2 AI auditing . . . 8
2.3 Insurance . . . 9
2.3.1 Health insurance . . . 9
2.3.2 Life insurance . . . 9
2.4 Synthesis . . . 10

**3 Method** **11**
3.1 Design science research . . . 11
3.1.1 Problem identification and motivation . . . 12
3.1.2 Objectives definition . . . 12
3.1.3 Design and development . . . 13
3.1.4 Demonstration . . . 14
3.1.5 Evaluation . . . 14
3.1.6 Communication . . . 15

**4 Regulations, standards and frameworks** **16**
4.1 Regulations . . . 16
4.1.1 AI Act . . . 16
4.1.2 GDPR . . . 21
4.1.3 DMA . . . 22
4.1.4 DSA . . . 22
4.2 Harmonized standards . . . 22
4.2.1 CEN/CENELEC JTC21 . . . 22
4.2.2 ISO/IEC . . . 23
4.2.3 IEEE . . . 24
4.3 Frameworks . . . 24
4.3.1 NIST . . . 25
4.3.2 NOREA . . . 26
4.3.3 KPMG . . . 26
4.3.4 IIA . . . 27
4.3.5 ForHumanity . . . 27

**5 Requirements and specifications** **29**
5.1 Literature review . . . 29

5.2 Exploratory interviews . . . 29
5.3 Contribution arguments . . . 31

**6 Synthetic data** **32**
6.1 Case construction . . . 32
6.1.1 Case: NexaSure . . . 32

**7 Design** **36**
7.1 Preparation process . . . 36
7.1.1 Scope of the audit . . . 38
7.1.2 Auditor requirements . . . 40
7.2 Audit controls . . . 40
7.2.1 Summary Audit controls . . . 42

**8 Evaluation** **43**
8.1 Evaluation interviews . . . 43
8.1.1 Relevance . . . 44
8.1.2 Usability . . . 44
8.1.3 Completeness . . . 44
8.1.4 Understandability . . . 45
8.1.5 Logic . . . 45
8.1.6 Added value . . . 45
8.1.7 Additional comments . . . 45

**9 Discussion** **47**
9.1 AI audit framework . . . 47
9.1.1 IT auditor expertise . . . 48
9.1.2 Standardization versus sector-specific customization . . . 48
9.1.3 Fairness, bias, and data drift . . . 49
9.1.4 The maturity of the AI auditing ecosystem . . . 49
9.2 Practical implications . . . 50
9.3 Limitations . . . 50

**10 Conclusions** **52**
10.1 Answers to the research questions . . . 52
10.2 Contributions . . . 53
10.3 Future work . . . 54

**A Interviews** **60**
A.1 Exploratory interview questions . . . 60
A.2 Evaluation interview questions . . . 61
A.3 Thematic coding . . . 61

**B Synthetic data** **64**

**C Audit controls** **73**

# Chapter 1

# Introduction

## 1.1 Problem statement

Due to the Artificial Intelligence Act (AIA) that is applied in the European Union (EU) and the growth of European investments in Artificial Intelligence (AI), a significant demand for AI assessments is currently emerging. The obligations that are stated for 'high-risk' systems in the AIA include different aspects, such as data governance, technical documentation, transparency, risk management and conformity assessments [64]. Even though the AIA doesn't implicitly state that audits are obligated, companies demonstrate efforts to abide with the AIA's requirements through compliance and readiness assessments, or assurance services. When companies do not comply to these obligations, they could receive strict penalties. Therefore, companies that use high-risk AI systems need these to be compliant to the provisions of the AIA, in order to receive an EU Declaration of Conformity [64].

To audit AI systems in the future, proper terminology, frameworks and transparency is needed. A harmonized framework for both internal and external audits of AI systems is necessary for this. However, this shared framework does not exist yet. This is partly due to the uncertainty of how the requirements of the AIA can be transformed to measurable criteria for an AI audit. The AIA stated that the EU would provide a guideline for providers of high-risk AI systems on the 2nd of February, 2026 [69]. However, the development of this guideline was still not finished on the 24th of February, 2026, which was already an extended deadline [33]. This guideline would be beneficial to providers of AI-driven risk assessment models used for pricing in health and life insurance, as this system falls under the high-risk category. Researchers stated that these models are frequently used by health and life insurance companies (HLIC) and need to be audited in the future [57, 55]. Nagamani [57] emphasizes that limited studies have been done about audit frameworks for explainable AI in health and life insurance.

For these reasons, the problem is that it is not known how AI-driven risk assessment models used in HLIC can be audited by a third-party audit firm under the AIA. Currently, no shared proper frameworks exist to audit such systems under the AIA, even though the need for these is likely to increase in the future. Shared proper AI audit frameworks could improve consistency between AI audits and increase trust in AI systems.

This research is very relevant to conduct since the world is becoming more dependent on AI, and this will only increase in the future. However, there have been incidents where AI is used in a negative way. AI algorithms have been used for social scoring decision-making models that turned out to be discriminatory [8]. This may also be the case in AI-driven risk assessment models used by HLIC [79]. Therefore, it is beneficial for society that these AI systems meet the requirements of Responsible AI (RAI). RAI refers to a collection of guidelines and frameworks designed to reduce the potential risks and challenges that can arise when using AI-driven systems [27].

Building on the concept of RAI, the EU introduced the AIA, which follows a risk-based approach with four categories. These are 'Minimal risk', 'Limited risk', 'High risk', and 'Unaccepted risk'. Provisions for AI systems that fall under the high-risk category will apply from the 2nd of August, 2026. To determine compliance with the AIA, readiness assessments are currently done by third-party audit firms. These assessments are not the same as audits, because assessments fall under advisory services. Audits are

more of a retrospective process (what happened), while assessments are more about what could happen in the future [49]. Since the AIA does not mandate audits, some providers of high-risk AI systems may not audit their AI systems, as the incentive is missing. However, this situation may change in the future and third-party audit firms are currently developing frameworks for this as well. Conformity assessments have to be conducted to prove compliance with the AIA. For risk assessment models used in HLIC, a conformity assessment can be performed internally. However, compliance audits conducted by third parties may be more trustworthy and credible due to their independence and critical perspective [7]. This provides the organizations with more confidence that their AI systems comply with the AIA and will receive an EU Declaration of Conformity. Therefore, the demand for audits by third-party audit firms is likely to increase, particularly among governmental institutions that place a high value on trust and reputation. Given the limited research on AI auditing, particularly the auditing of AI-driven risk assessment models, this study makes an important contribution to the existing academic literature.

## 1.2 Research questions

The first objective of this research is to investigate how AI systems are currently assessed by third-party audit firms. The second objective is to investigate what specifications and requirements third-party auditors should apply to audit AI-driven risk assessment models used in HLIC in the future, due to the AIA. The third objective is to improve an AI audit framework used by third-party auditors, so that it could assist them in auditing AI-driven risk assessment models used in HLIC, under the AIA. This will then be tested on a mock demonstrator case. The fourth objective is to evaluate the upgraded audit framework through an assessment conducted by third-party auditors. From these four objectives, the following research question is derived.

**RQ.** *How can an AI audit framework used by a third-party audit firm be improved to effectively audit AI-driven risk assessment models used in health and life insurance under the AI Act?*

### 1.2.1 Sub-questions

This research question can be divided in to 4 sub-questions. Each sub-question gives an answer to one of the objectives, as these also match with each other. These answered combined will provide the answer to the main question. The sub-questions are stated as follows:

**SQ.1** *What mechanisms do third-party audit firms currently use to assess AI systems?*

To answer the main research question, we first need to understand the current situation within third-party audit firms. By analyzing these mechanisms, we can identify their capabilities and limitations, which helps us improve the AI audit framework.

**SQ.2** *Following the adoption of the AI Act, what specifications and requirements must third-party auditors apply when auditing AI-driven risk assessment models used in health and life insurance?*

This question focuses on the specifications and requirements for the AI audit framework that arise from the adoption of the AIA. These specifications and requirements are relevant to the research, as they determine how we can improve the AI audit framework to audit AI systems under the AIA.

**SQ.3** *To what extent can an upgraded AI audit framework fulfill these specifications and requirements when tested through a mock demonstrator case?*

For this question, we developed a mock demonstrator case in which we tested the AI audit framework on a fictive scenario. By doing so, we were able to illustrate how an AI audit under the AIA can be conducted using the AI audit framework.

**SQ.4** *What is the assessment of third-party auditors regarding the upgraded AI audit framework, as illustrated through a mock demonstrator case?*

The final sub-question provides insights into the quality and usability of the AI audit framework. We achieved this by having third-party auditors evaluate the upgraded AI audit framework, when it is illustrated through a mock demonstrator case. After all the sub-questions are answered, we provide recommendations for the auditing processes of AI systems. These recommendations are based on the findings and reasoning of the answered sub-questions.

## 1.3 Overview of the thesis

In Chapter 2, we discuss the background and related work that is relevant to this study. Topics regarding artificial intelligence are included here, as well as information about health and life insurance. We present the method of the research in Chapter 3 and we discuss the literature review in Chapter 4. Subsequently, the requirements and specifications for the AI audit framework are described in Chapter 5. Then, we present the synthetic data for the mock demonstrator case in Chapter 6. In Chapter 7, we present the design activities of the audit framework and explain why we made certain choices. Finally, we present the discussion in Chapter 9 and the conclusions in Chapter 10.

# Chapter 2

# Background and related work

The background and related work chapter consist of existing academic literature that is relevant to our research. First, we elaborate on AI systems, model metrics and RAI. Then, we discuss the current field of AI auditing. Finally, we explain how underwriters in health and life insurance operate, with a focus on the Netherlands.

## 2.1 Artificial intelligence

To create a better understanding of how artificial intelligence (AI) systems can be audited, we elaborate on this technique in this section. AI is a broad term that contains any form of intelligent behavior by algorithms [75]. This can vary from algorithms that can play chess to algorithms that recognize objects in pictures. To put it in simple words, AI receives data as an input and produces predictions as an output. Machine learning (ML) is the most used AI technology. This technology is able to learn from training data by correcting the relations between input data and the target variable, so that predictions become more accurate. For our research, we focus on machine learning models, as these are primarily used for risk assessment in health and life insurance [55].

### 2.1.1 Black box model

A black box model refers to a model from which its internal logic is not known or not interpretable [31, 32]. Machine learning models could contain learning processes which are hard to understand, because they are too complex for humans. Therefore, the explainability and transparency of such AI models are lacking. Predictions that are made from a black box model are hard to justify, which is crucial to do in critical sectors like banking and healthcare, for example. Furthermore, it is hard to improve the accuracy of a black box model, due to the incomprehensible learning processes of this AI model.

### 2.1.2 AI-driven risk assessment model

Onwuajuese et al. [70] state that an AI-driven risk assessment model can be defined as an AI algorithm that can use large datasets to estimate risks. Machine learning techniques enable the ability to discover patterns and potential risks that are complicated to find manually. AI-driven risk assessment models use real-time data and need less human input than traditional risk assessment models. There are different sectors that use AI-driven risk assessment models for different purposes. For instance, AI-driven risk assessment models are frequently used in weather forecasting to predict typhoons, floods and droughts by analyzing real-time meteorological data [52, 19]. Nagamani [57] studied the uprising use of AI-driven risk assessment models in HLIC. The paper states that these models are used to calculate the odds that an applicant will file a claim. Based on this, the model estimates the risk score, which determines the premium price. This process is also known as underwriting. Furthermore, Nagamani explains how these models can help to improve risk assessment accuracy and reduce underwriting time. On the other hand, it describes the difficulties that come with this, regarding regulatory compliance and ethics. Mitta [55] examines how different AI techniques are used for underwriting and how these techniques can be combined to one model. This paper also addresses the challenges regarding biased data, transparency, explainability and privacy. Therefore, it emphasizes the urgency to perform regular audits on these

AI-driven risk assessment models. Although such models are applied in many domains and use cases, our thesis focuses only on AI-driven risk assessment models used for underwriting in health and life insurance.

**Model metrics**

To measure the performance of an AI-driven risk assessment model, different metrics can be used. Models that produce risk scores from 0 to 1 are defined as regression models. These models can be measured, among other metrics, with the $R^2$ method. The $R^2$ defines how much variance of the data can be explained by the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(\bar{Y} - Y_i)^2} \tag{2.1}$$

The formula of this method is shown in eq. (2.1). Here, $X_i$ means the predicted score by the model and $Y_i$ means the actual score. The average value of all $Y_i$ values is defined as $\bar{Y}$. The $\bar{Y} - Y_i$ part calculates the total variance of the data ($SS_{tot}$). The $X_i - Y_i$ part calculates the variance that is not explained by the model ($SS_{res}$). By expressing $SS_{res}$ as a proportion of $SS_{tot}$ and subtracting this value from 1, the proportion of variance explained by the model is obtained [13, 81].

Another metric that can be used to measure the performance of a regression model is the Root Mean Square Error (RMSE). This method measures the difference between the predicted scores and the actual values of a model.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2} \tag{2.2}$$

The formula of RMSE is shown in eq. (2.2). With this method, the high outliers are 'punished' more relatively, by squaring the difference [81]. Furthermore, the Population Stability Index (PSI) can be used to measure the differences between population data over time. This is relevant for risk assessment models used by HLIC, as these models are often trained on historical data and need to produce accurate risk assessments when applied to new data [84]. If the differences in population data are relatively large, this is referred to as data drift. Mannapur [54] explains that this phenomenon occurs when the relationship between the input data and the target variable differs from what it was previously. Therefore, the predictions of the model cannot be trusted anymore. Mannapur also explains that data drift is a main risk in the healthcare sector, as the health of a society can change over time. This data drift then impacts the accuracy of patient risk assessments in hospitals.

$$PSI(Y_b, Y; B) = \sum_{i=1}^{B}(y_i - y_{b_i})\big(\ln(y_i) - \ln(y_{b_i})\big) \tag{2.3}$$

The formula of the PSI is shown in eq. (2.3). The PSI is measured by grouping scores into categories, also called bins, using a multinomial classification approach. The PSI compares the proportions of the target year, which is referred to as $y_1, ..., y_B$ , with the base year, which is referred to as $y_{b1}, ..., y_{bB}$. The $i$ refers here to the index of the bin, where $b$ stands for the amount of base and $B$ for the amount of bins [84].

### 2.1.3 Responsible AI

Many academic researchers have discussed what Responsible AI (RAI) encompasses. According to Goellner et al. [29], who conducted a literature review on RAI, RAI is a human-centered approach to the development and use of AI systems. It aims to ensure trustworthy outcomes through ethical and explainable AI systems, while principles such as privacy and security are also important.

According to Gadekallu et al. [27], RAI is a set of frameworks and principles to mitigate the challenges and risks that may occur in AI-driven applications. They explain that the main principles of RAI are transparency, accountability, fairness, privacy, robustness, and safety. Gadekallu et al. [27] also emphasize the urge to implement RAI into today's applications. However, this goal faces some challenges due to several reasons. RAI applications often provide less profit than regular AI applications, which makes it less ideal for organizations to implement RAI. Furthermore, the current regulations are not strong enough

to encourage organizations to implement RAI. Guidelines by authorities for the implementation of RAI are often abstract and multi-interpretable. There is a lack of RAI expertise within most organizations, thus training is needed to recognize the RAI principles within AI applications [27].

## 2.2 AI auditing

An audit can be defined as a "systematic, independent and documented process for obtaining objective evidence and evaluating it objectively to determine the extent to which the audit criteria are fulfilled", according to the ISO 19011:2018 [37]. Audits can be described as assurance engagements, where the purpose is to verify compliance with standard practices in a specific domain. This is something different than an assessment, which is an advisory engagement that evaluates maturity, readiness, or future risks, intended to add value and improve the governance, risk management, and control processes of an organization [49].

First, Koshiyama et al. [41] describe AI auditing as a set of approaches to monitor, assess, and assure the safety, compliance, and ethics of AI and associated algorithms. This is done by embedding appropriate socio-technical controls to manage and mitigate risks that may occur regarding the AI system. Audits may be conducted either internally by the organization itself or externally by independent or interested parties. Audits performed by independent audit firms are commonly referred to as third-party audits. This distinction matters for this research, as we focus here on audits performed by third-party audit firms. Mökander [56] describes some challenges for the AI auditing processes, such as the lack of shared terminology and standards for AI auditing. This leads to confusion and inconsistent practices. Furthermore, there are limited frameworks for AI auditing and the effectiveness of the frameworks that exist is not studied.

Li et al. [50] provided the first comprehensive mapping of AI auditing and auditability research. They concluded that AI audits are vital to create responsible AI, but effective auditing depends on building auditability into systems from the start. The study emphasizes that documentation, standardization, and multi-stakeholder collaboration are key enablers for transparent and accountable AI governance. Novelli et al. [62] examines how the AIA can be effectively implemented and enforced through a coordinated governance structure. It outlines an institutional framework that is necessary for regulating AI across the EU, focusing on both supranational and national levels.

Raji et al. [74] explain that external AI auditing is limited due to the lack of sources that auditors receive from the client, also known as evidence. Documents about internal processes and training data are often strictly confidential and therefore not shared with the external auditor. Because of this, the auditability of AI systems is less sufficient for external audits than for internal audits. Multiple papers emphasize the challenge of operationalizing the AIA's abstract legal requirements into concrete technical measures [63, 78]. Ntaoulas et al. [63] are currently developing an audit framework for AI systems under the AIA, where they try to transform the regulatory requirements into measurable criteria. Sarkar et al. [78] explain the need for clear methodologies to ensure compliance and effective auditing of AI systems. Translating 'high-risk' principles such as 'appropriate levels of accuracy' or 'effective human oversight' into measurable and verifiable criteria requires significant technical and organizational expertise that many organizations may lack. Lam et al. [48] introduced a model for conducting AI audits aimed at ensuring compliance and building assurance. They emphasized that independent auditors play an essential role in confirming that AI systems are responsibly and reliably designed, developed, implemented, and overseen. According to their analysis, effective AI audits should include four essential elements:

- Audits must rely on a standardized and publicly accessible set of evaluation criteria.
- The purpose of the audit is to check compliance or provide assurance by assessing the system against an established normative framework, such as the AIA.
- The quality of AI auditing is crucial, meaning AI auditors must be trained properly.
- The audit reports should be made publicly available, at least partially, to enhance transparency and accountability.

In addition to this, several aspects are recommended to use when performing an AI audit. These originally stem from financial audit processes and are as follows:

- Define the audit scope.
- Gather and review documentation about the AI system and the processes regarding the use of AI. Check whether these documents meet the requirements for the audit.
- Validate the evidence of the documentation. These may include interviews with the client, examining the bias testing of the AI system, and checking the data protection impact assessment.
- The publication of the audit report should provide the findings in a publicly shared format and should include the assessment of the auditor.
- In the end, a certificate is provided to show how well the AI system complies with the specific regulation or standard.

## 2.3 Insurance

Individuals can protect themselves against a wide range of life risks by purchasing insurance. Insurance companies offer coverage for property, services, and human lives, for example. The goal of these companies is to operate profitably. Therefore, insurers assess and predict the likelihood that an applicant will file a claim in the future. This underwriting process, as mentioned in Section 2.1.2, is an assessment that contains data-driven analyses. Underwriters estimate the risks and determine appropriate premiums and policy conditions based on these risks [76].

### 2.3.1 Health insurance

To cover the costs of medical care, individuals can purchase a health insurance [76]. Costs related to hospitals, physiotherapists, dentists and pharmacies can be fully or partially covered by the insurer. In some EU countries, purchasing a health insurance is mandatory. In these countries, such as the Netherlands, individuals are required to obtain a basic health insurance policy and may choose to purchase supplementary coverage to insure additional medical services. In the Netherlands, insurers are legally prohibited from rejecting applicants for the basic health insurance policy and from performing underwriting to determine the premium for this coverage [3]. However, underwriting is permitted for the supplementary health insurance, where insurers are allowed to assess risk and differentiate premiums. For this reason, AI-driven risk assessment models may only be used for the underwriting process in the supplementary health insurance segment. Health insurance typically results in large amounts of reported claims, as nearly everyone requires medical care at some point in their life. Therefore, it is crucial for health insurers to carry out their underwriting processes as accurately and efficiently as possible in order to control costs due to claims. Therefore, the underwriters can use personal, medical, and health data that is crucial for the underwriting process, if the applicant provided this or gave consent for this. This is one of the obligations of the GDPR, which is explained further on this thesis. The data that insurers ask from the applicants for underwriting differs across insurance companies [80].

### 2.3.2 Life insurance

Individuals can protect themselves against the financial consequences of premature death by purchasing life insurance [76]. In the event of an early death, the insurer pays out a benefit that can help surviving family members cover expenses such as living costs, outstanding debts, or funeral costs. Purchasing a life insurance is voluntary and individuals can decide for themselves when to purchase this. The premium paid for life insurance is influenced by factors such as age and lifestyle at the start of the policy. Younger individuals with a healthy lifestyle typically pay lower premiums than older individuals or those with less healthy lifestyles. These factors have an impact on the life expectancy of the individual, which is important to the insurer. In the Netherlands, the insurer collects data about the applicant's age and whether he or she smokes. If the insurer believes additional information is needed about the applicant's medical history for underwriting purposes, it can process this after the applicant provided this or gave consent for this, due to the GDPR. Similarly, in life insurance, the data used for underwriting varies across insurers [3, 80]. This makes it complicated to design a standardized audit framework for AI-driven risk assessment models used by HLIC, as the input data for the model influences the sort of risks and biases that can be introduced.

## 2.4 Synthesis

This Chapter showed that AI-driven risk assessment models are increasingly used in health and life insurance to support underwriters. However, this introduces challenges related to transparency, explainability, fairness, bias, privacy, and data drift, because of the use of personal data by underwriters. Existing literature highlights the importance of RAI and regular auditing of AI systems. Nevertheless, most organizations lack sufficient AI expertise to audit AI systems, as many different AI techniques exist and black box models are relatively complicated to understand. AI auditing remains an emerging field that lacks shared terminology and standardized audit procedures. Researchers argue that limited research has been done on the auditing of AI-driven risk assessment models used in health and life insurance under the AIA. Existing frameworks mainly provide abstract principles rather than concrete audit controls that can be applied by third-party auditors. Therefore, the research gap is that there is currently no validated and structured audit framework that enables third-party auditors to effectively audit AI-driven risk assessment models used by HLIC under the AIA.

# Chapter 3

# Method

## 3.1 Design science research

This chapter explains the methodological approach that is used for this study. With this approach, we investigated how an AI audit framework used by a third-party audit firm can be improved to effectively audit AI-driven risk assessment models used by HLIC under the AIA. This study is conducted with the method of design science research [72].

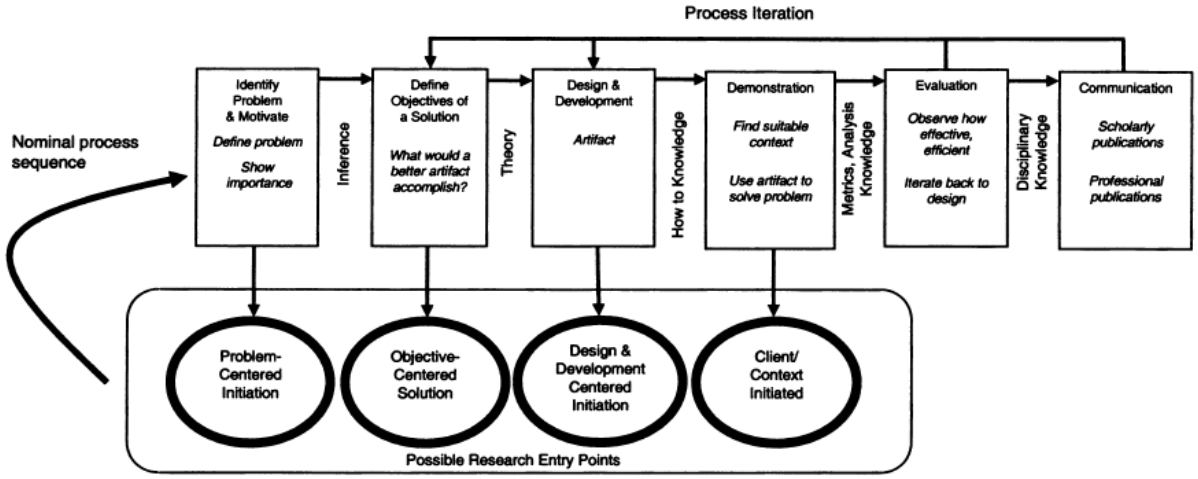



Figure 3.1: The six phases of design science research illustrated [72].

K. Peffers et al. demonstrate how this research methodology can be used in the field of information systems. Design science research is characterized by the fact that an artefact is developed during the research. Considering that the main objective of this thesis is to design an AI audit framework, we chose this methodology for our research. Figure 3.1 illustrates the six phases of design science research. These phases are Problem Identification and Motivation, Definition of the Objectives for a Solution, Design and Development, Demonstration, Evaluation, and Communication [72].

Furthermore, this research is conducted by doing a case study. Breslin et al. [10] explain that a case study method can be used for research, especially in design science research. Both research methods are a proper way to explore the world of theory combined with the practical world. Theoretical frameworks made with design science research can be tested and evaluated with a case study. This provides the theoretical development with empirical grounding. Breslin et al. [10] refers to this concept as a design case study. This approach uses a case study that is embedded within the design science research methodology. Here, the case study is related to the IT Assurance & Advisory (ITAA) department at KPMG Nederland, which is the third-party audit firm where the internship is followed. KPMG is one of the biggest third-party audit firms in world and is part of the 'Big Four' consultancy firms. In the last decades, KPMG has built up extensive expertise in IT assurance and advisory [45]. Therefore, the ITAA department at KPMG is a

suitable and credible environment in which to conduct a case study. The case study includes an AI audit of a fictive AI-driven risk assessment model that is used in a fictive health and life insurance company. We tested the framework by asking experts to audit the fictive AI-driven risk assessment model with the designed audit framework. This is referred to as a mock demonstrator case. Since the experiment is conducted within KPMG, we were only able to draw final conclusions about KPMG's practices. In this section, the research method will be further explained by elaborating on the case study method and the six phases of the design science research, which are integrated into each other.

### 3.1.1 Problem identification and motivation

The problem identified in this thesis, was earlier explained in Section 1.1. The problem is that it is not known how AI-driven risk assessment models used by HLIC can be audited by a third-party auditors under the AIA, as shared audit frameworks and guidelines do not exist. Non-compliance with the provisions of the AIA by providers of high-risk AI systems will result in penalties imposed by the EU. Besides the reputation damage that providers may suffer, their AI system's decisions could also have a big impact on people's daily lives. For instance, AI systems used in financial markets, hiring, and law enforcement could make predictions about people that may influence their life significantly. Such outcomes would be particularly concerning if they result in detrimental effects for specific subgroups of society. Therefore, conducting this research is highly relevant.

### 3.1.2 Objectives definition

The objectives of this research are derived from the problem statement. These objectives are defined in Section 1.2. Wieringa et al. [82] explains that there are different goals when using the research science design method.

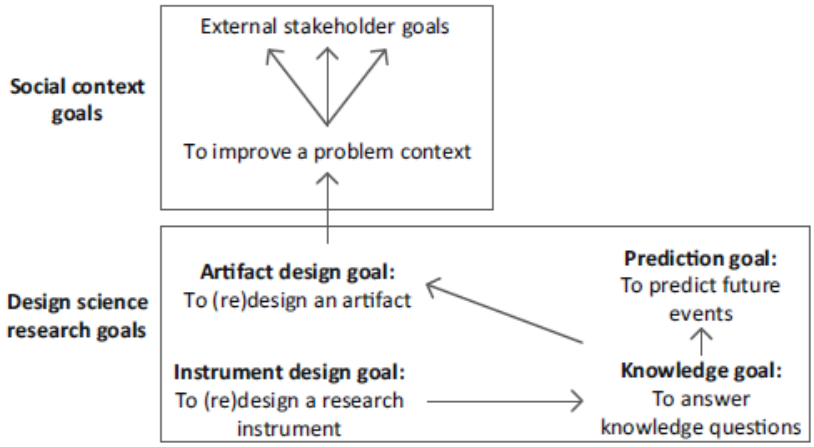



Figure 3.2: A visualization of how different goals within research science design relate to each other.

Figure 3.2 illustrate how different goals within research science design relate to each other. First, researchers have prediction goals, which is the desire to know how a designed artifact would operate in a certain context for example. For our research, we want want to know how AI systems could be audited under the AIA. Furthermore, we have knowledge goals that describe how we could know if this prediction will be true. For our research, the knowledge goal is to know what happens when our AI audit framework is tested on a mock demonstrator case. To answer this question, we need to know how to design the instrument or artifact. Therefore, instrument design goals are required for the design of our AI audit framework. Our instrument design goal is to know what requirements and specifications are needed to build the AI audit framework.

Moreover, we have an artifact design goal, which is to improve the functionality of our AI audit framework within a context. This often supports external stakeholder goals, which are based on different needs. In this context, these stakeholders are IT auditors and clients of KPMG, regulators, and natural persons that are subject to the use of AI systems. For our research, we obtained the instrument design goal and

artifact design goal via the exploratory and evaluation interviews with experts of KPMG, respectively. Moreover, we gathered the stakeholder goals of regulators, which is the EU, by analyzing the applied regulations. By identifying these goals, it became clear how the AI audit framework could be designed and improved over time. The relations between these goals are illustrated in figure 3.2.

### 3.1.3 Design and development

This phase consists of the design and development of the AI audit framework for the auditing of AI-driven risk assessment models used in HLIC. To define the requirements and specifications for this framework, a literature review and exploratory interviews are conducted.

#### Literature review

For the literature review, existing frameworks and standards for AI auditing are studied, together with the relevant EU regulations. Besides this, academic articles regarding AI auditing and risk assessment models used in HLIC are studied. Thus, the literature review consists of academic literature and grey literature [28]. Grey literature can be online blogs, magazines, technical reports or insights from companies for example. This can be very useful for our study, since the topic is very recent and specific developments can occur at any time. Academic literature is unable to keep pace with this rapidly changing environment, whereas grey literature provides more up-to-date work and sources.
After this, we translated the findings of the literature review to the requirements and specifications for the AI audit framework for AI-driven risk assessment models used by HLIC, under the AIA.

#### Exploratory interviews

Moreover, we needed in-depth and qualitative insights from experts to define the requirements and specifications for the AI audit framework. To obtain these insights, we conducted semi-structured and exploratory interviews with five employees from the IT Audit & Advisory department at KPMG. Some interviewees are part of the 'RAI department', which is a small team that focuses on RAI projects and is part of ITAA. We chose this amount of interviewees to allow the interviewer to conduct long, in-depth interviews and the opportunity to ask follow-up questions. Furthermore, considering that a second round of interviews would be conducted at a later stage, we deemed five interviewees to be feasible within the project timeline. Each interview lasted around 30 minutes.

| | Role | Experience | Expertise |
|---|---|---|---|
| Expert 1 | Junior consultant | ± 1 year | Responsible AI |
| Expert 2 | Junior consultant | ± 1 year | Responsible AI |
| Expert 3 | Manager | ± 7 years | Responsible AI |
| Expert 4 | Director | ± 12,5 years | IT audit, DSA, DMA |
| Expert 5 | Partner | ± 15 years | IT audit, DSA, DMA |

Table 3.1: The profiles of the interviewees of the exploratory interviews.

The profiles of these experts are presented in table 3.1. We selected the interviewees based on their professional skills and experience in IT auditing, RAI, and relevant regulations and standards. Furthermore, we selected interviewees from a diverse range of seniority levels. By choosing the interviewees this way, we aimed to capture a variety of perspectives. We expected that this diversity facilitates a broader range of responses from different perspectives, as it is likely that someone with a RAI perspective has different opinions about AI auditing than someone with an IT auditing perspective. Their experiences provide valuable perspectives on both the practical and strategic dimensions of AI assurance. For these reasons, we believed that five interviews were sufficient to gather the required insights to answer the research questions.

The interviews aimed to explore how these experts currently approach AI auditing, including the methods and frameworks used within ITAA. Topics included the perceived effectiveness of current audit practices, key challenges in assessing AI systems, and the organizational or regulatory changes they believe are necessary to stimulate a well-functioning AI auditing ecosystem. Finally, we asked the interviewees what they think is extra necessary to audit on when auditing AI-driven risk assessment models used by

HLIC. The interview questions can be found in Appendix A.1. We transcribed the interviews verbatim, after which we systematically identified and categorized key insights using open coding and thematic coding. The purpose of this phase was to translate theoretical insights and practical expert perspectives into a structured set of criteria, steps, and considerations that auditors can use when auditing high-risk AI systems. In the end, we translated the insights derived from each code into requirements and specifications for the AI audit framework.

### 3.1.4 Demonstration

The design of an artifact does not resolve a problem. The interaction between the artifact and the context of the problem could help in resolving the problem [82]. Therefore, we tested the designed audit framework through a mock demonstrator case. A mock demonstrator case is a fictive yet realistic example that can be used to apply a designed artefact to. The idea was to simulate an AI audit of a fictive AI-driven risk assessment model by applying it to the upgraded audit framework. Within this process, different parts of the upgraded audit framework were tested.

This mock demonstrator case is about a fictive AI development company, which is called NexaSure. NexaSure develops AI-driven applications, such as risk assessment models that predicts the risk of a new health or life insurance contract with an applicant. This risk score is then used to set the price for the applicant. AI systems that are used for this process fall under the high-risk category of the AIA, which is why NexaSure wants to let KPMG audit its AI system to ensure compliance with the AIA. Because the internship was followed at KPMG, access has been granted to confidential documents and business processes of KPMG. Using these documents, we were able to construct a realistic mock demonstrator case based on prior audits. These documents included audit methods and reports, client lists, and the procedures of gathering evidence. After we constructed the mock demonstrator case, an IT auditor from KPMG assessed the case to ensure it accurately reflects real-world conditions.

### 3.1.5 Evaluation

For the evaluation process, we interviewed five experts from KPMG again. The testing of the upgraded framework through the mock demonstrator case was evaluated during these expert interviews. The experts were asked to reflect on the mock demonstrator case and provide feedback on the framework. This way, the practical usability of the AI audit framework was assessed. The framework was evaluated on its effectiveness in supporting the auditing of AI-driven risk assessment models used by HLIC, under the AIA. To ensure validity and objectivity, the evaluation interviews were conducted separately. The quality criteria of Lincoln et al. [51] was used to ensure the trustworthiness of the evaluation interviews [51]. These criteria are explained as follows:

- **Credibility**. This means that we need to verify whether our interpretations align with the interviewee's experiences and meanings. This was achieved by asking clarifying and control questions during the interviews.
- **Transferability**. This means that we should gather insights regarding to contextual boundaries of the results. This was achieved by asking questions about the usability of the framework.
- **Dependability**. This refers to the trustworthiness of the evaluation interviews. The data from the interviews are trustworthy if all the interviews follow a strict semi-structured interview protocol.
- **Confirmability**. This means that the conclusions that will be drawn from the interviews come from the data, and not from our own assumptions. This was achieved by asking the interviewees to justify their judgments with explanations and by grounding conclusions in direct quotes.

The interview questions included six topics, which are relevance, usability, completeness, understandability, logic, and added value. These topics were derived from the quality criteria of Lincoln et al. [51]. We asked the interviewees to rate the framework on each topic, on a scale ranging from 1 to 10. We chose this rating scale because Joshi et al. [47] argue that increasing the amount of rating scales generally improves validity and reliability of the respondents' evaluations. Furthermore, the authors of this paper emphasize that open questionnaires are more effective than closed questionnaires, as they allow interviewees to explain their answer in detail. For this reason, we asked the interviewees to explain why they gave a certain rating. The questions for the evaluation interviews can be found in Appendix A.2. After

conducting the evaluation interviews, we transcribed the interviews verbatim and assessed the ratings of the interviewees with descriptive statistics.

### 3.1.6 Communication

In the end, this research and the end product is communicated to the Leiden University and to KPMG. This will be done with a presentation about the research and the artefact that is developed. Finally, the artefact will be provided to KPMG.

# Chapter 4

# Regulations, standards and frameworks

In this chapter, we elaborate on the different regulations, standards and frameworks that are applied to regulate and assess AI and data usage, with a focus on the EU. These three concepts are relevant for this research, as they determine the requirements and specifications for the upgraded AI audit framework. The regulations contain the legal requirements, which we have to acquire to know what AI-driven risk assessment models used by HLIC have to comply with. The existing standards and frameworks for assessing and auditing AI systems provide us with knowledge about the current field of AI auditing by third-party audit firms. Based on these, we determine which AI principles and audit methods are relevant for the upgraded AI audit framework.

## 4.1 Regulations

This section describes the relevant EU regulations that organizations located within the EU and organizations from outside the EU that operate within the EU have to comply with. Many researchers emphasize the importance of the development of regulations regarding AI, data and privacy [53, 27]. They conclude that too little regulatory structures exist in this field, even though the public awareness about the negative aspects of AI is growing. There are many stakeholders with different priorities, such as engineers, researchers, policymakers, and clients. These people need to work together and discuss what is important, in order to develop proper regulations that are well accepted by everyone.

### 4.1.1 AI Act

Drawing on the foundations of RAI, which we described in Section 2.1.3, the EU created the AI Act (AIA). The AIA entered into force on the 1st of August, 2024 [24]. With this act, the EU seeks to create a fair and safe 'AI ecosystem', from which every working sector and every citizen will benefit. The term AI ecosystem refers to all the development and deployment of AI systems. The AIA defines an 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" [69]. AI systems that are only used for research purposes, or for military or national security purposes, are excluded from the AIA.

The AIA was first proposed by the EU Commission in April 2021 and approved by the European Parliament and the Council in December 2023. The AIA is designed to protect civilian's health, safety, and fundamental rights from potential AI-related risks. The AIA sets out clear rules and responsibilities for those who build and use AI systems, while also aiming to minimize the administrative and financial burden on businesses. The EU Commission explains that these regulations were needed because some AI systems may bring risks for certain civilians that harm their health, safety, or fundamental rights [18]. For instance, when an AI system cannot explain why it makes specific predictions, this could harm the fundamental rights of individuals. When these predictions lead to a disadvantage for a person in a critical context, it is hard to determine whether this is fair or not.

Article 3 of the AIA defines several roles and terms that are crucial within the regulation [69]. Some of

these are defined in the AIA definition section as follows:

- **Provider**: "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" [69]. This is the most relevant role for our research, as we focus on providers when developing the upgraded AI audit framework.
- **Deployer**: "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" [69].
- **Authorized representative**: "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" [69].
- **Importer**: "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" [69].
- **Distributor**: "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" [69].
- **Operator**: "A provider, product manufacturer, deployer, authorized representative, importer or distributor of an AI system." [69]
- **Conformity assessment**: "The process of demonstrating whether the requirements set out in Chapter III, Section 2 relating to a high-risk AI system have been fulfilled" [69].
- **Notified Body**: "A body that performs third-party conformity assessment activities, including testing, certification and inspection. This body is notified in accordance with this AIA and other relevant Union harmonization legislation" [69].

The AIA has a risk-based approach, which means that it indicates different obligations to AI systems based on their risk [18]. The AIA follows the traditional view that risk refers to the probability that a potential hazard will actually result in harm, loss, or damage [61]. An increase in safeguards will lead to a decrease of the likelihood that a hazard occurs.

$$risk = \frac{hazard}{safeguards} \tag{4.1}$$

The expression of risk as the ratio between hazard and lifeguards is shown in eq. (4.1) [61]. Normally, risk tolerance thresholds determine which risks are still accepted, and which risks need mitigation practices or even are prohibited. In the AIA, these thresholds form the four risk categories [61]. This is based on whether an AI system could potentially harm the safety, health or fundamental rights of EU civilians. The AIA categorizes AI systems into four different categories. These categories are as follows [69]:

- **Minimal risk**. AI systems that fall under this category contain (almost) zero risks. The majority of all AI systems used in the EU fall under this category. Examples of this kind of systems are spam filters and AI-enabled games.
- **Limited risk**. This category contains AI systems that have limited risks. These systems are referred to as General-Purpose AI models (GPAI models). Systems such as chatbots and deepfakes will fall under this category. The only provision that providers of GPAI models need to comply with is about transparency. This includes that end-users should be aware that they are interacting with AI.
- **High-risk**. AI systems that fall under this category contain a significant amount of risks. Therefore, this category consists the most obligations, of which most are assigned to the provider of the system. These provisions will be explained further on.
- **Unaccepted risk**. The AI systems that contain an unacceptable risk are prohibited for development and deployment, according to the AIA.

### High-risk category

The high-risk category is the most relevant category for this research, because the AI-driven risk assessment models used by HLIC will fall under this category. Therefore, the high-risk category is more elaborated in this subsection, with a focus on the parts that are relevant for providers of AI-driven risk assessment models used in health and life insurance. The focus on providers is due to the reason that most of the AIA obligations are assigned to providers, instead of deployers, importers, distributors, or authorized representatives.

#### Classification of AI systems as high-risk

According to Article 6, Annex I and Annex III of the AIA [69], AI systems are defined as high-risk AI systems if they are used in one of the following fields:

- Safety components of products
- Biometrics
- Critical infrastructure (physical, digital, or utilities)
- Education and vocational training
- (Self-)employment and workers management
- Essential private services and essential public services and benefits
- Law enforcement
- Migration, asylum and border control management
- Administration of justice and democratic processes

However, the AIA also states that these AI systems do not fall under the high-risk category if they do not pose a significant risk to health, safety and fundamental rights. Nevertheless, if an AI system stated in the Annex III performs profiling of individuals, the AI system will always be considered as high-risk. Models used for risk assessment and pricing in relation to natural persons within health and life insurance fall under the category of 'Essential private services and essential public services and benefits' [69].

#### Risk management system

According to Article 9 of the AIA, providers of high-risk AI systems are required to establish a risk management system (RMS). The RMS should be able to recognize and assess risks related to the use of the AI system. Furthermore, it should implement measures to minimize these AI-related risks [69]. This could be complicated for the auditor to audit, as it is hard to determine what specific risks could be introduced by the use of the AI system in a specific sector.

#### Data and data governance

According to Article 10 of the AIA, models of high-risk AI systems should be developed with training, validating and testing datasets that meet certain criteria [69]. The datasets used for training, validating, and testing must follow data governance and management procedures that align with the intended purpose of the high-risk AI system. These procedures include clear design choices, bias identification and mitigation, and appropriate data collection and preparation procedures. The training, validation, and testing datasets must be relevant and sufficiently representative. Moreover, the data must be accurate and complete for their intended use, for as much as possible. Providers also need to possess the right statistical properties, including a reflection of the individuals or groups for who the high-risk AI system is designed [69].

Furthermore, the datasets must reflect the specific geographical, contextual, behavioral, or functional environment in which the high-risk AI system will operate. If it is necessary to detect or correct biases in datasets, providers may process special categories of personal data, such as sensitive data. However, this can only happen under strict safeguards to protect fundamental rights. Lastly, all previous mentioned requirements regarding data and data governance only apply to the testing datasets used for that AI system, if the provider does not train the model of the system itself [69]. To determine compliance with these requirements, third-party auditors must know what data is allowed to be used to train the model with and which biases can be introduced within the model.

**Technical documentation**
Article 11 establishes the rules for preparing and maintaining technical documentation for high-risk AI systems [69]. The purpose of this is to ensure that authorities can assess whether an AI system complies with all the requirements of the AIA, as all the evidence for the AIA obligations must be documented in here. Furthermore, it should include descriptions of the AI system's purpose, interactions, hardware and software details, development processes, and the used approach for compliance.

**Record-keeping**
According to Article 12 of the AIA, high-risk AI systems should be able to record events, also known as logging [69]. In order to ensure proper traceability of the AI system's functionality, logging processes should enable the recording of actions relevant for:

- Detecting situations that could lead to a substantial change or a risk regarding the health, safety or fundamental rights of the intended user, caused by the AI system.
- Facilitating the post-market monitoring process.
- Monitoring the performance of the AI system.

**Transparency and provision of information to deployers**
Article 13 states that high-risk AI systems must be designed and developed so that their functioning is transparent enough for deployers to understand the system's outputs and use them correctly [69]. High-risk AI systems must be supplied with user instructions, provided in a proper digital or other format. These instructions must be precise, complete, accurate, and easy for deployers to access and understand. First, this information should include details about the provider, including its contact details. Second, it should include details about the AI system's properties, and the qualities and limitations regarding system performance. Furthermore, it should include information about the implemented modifications, human oversight measures, hardware sources, expected lifespan, and instructions for record-keeping [69]. This is complicated to audit for third-party auditors, as the auditor should determine whether the user to able to understand the AI system, while the AI expertise of this user is unknown.

**Human oversight**
According to Article 14 of the AIA, high-risk AI systems must be created and developed so that natural persons can effectively supervise them while they are being used [69]. This should include proper human–machine interface tools for support. Human oversight should aim to prevent or reduce risks to health, safety, or fundamental rights that may arise when a high-risk AI system is used as intended or under reasonably foreseeable misuse. The measures must match the level of risk, the degree of autonomy, and the context in which the high-risk AI system is used. Individuals that use the AI system should be able to fully understand the AI system, decide to ignore the system's outcomes, and step in and take control of the AI system. Lastly, the user should be made conscious of automation bias, which is the risk of unconsciously trusting the outcomes of the model.

**Accuracy, robustness and cybersecurity**
Article 15 of the AIA states that high-risk AI systems must be created so that they perform properly on accuracy, robustness and cybersecurity aspect [69]. High-risk AI systems must be built to reduce errors, malfunctions, and inconsistencies that may arise, for as much as possible. To achieve this, organizations must implement both technical and organizational safeguards. For high-risk AI systems that continue learning after deployment, the design must ensure that biased outputs do not influence future inputs in a way that creates harmful feedback loops [69]. This is complicated for auditors to assess, as it is hard to determine when the accuracy of the model is sufficient.

High-risk AI systems must be capable of withstanding attempts by unauthorized third parties to exploit vulnerabilities in order to change the system's intended use, outcomes, or functionality. The cybersecurity measures put in place must be suitable for the specific context and the risks involved. The technical measures should include mechanisms to avoid, identify, react to, mitigate, and control attacks [69].

**Other obligations**
Beside the previous mentioned obligations, providers of high-risk AI systems must establish and maintain a quality management system (GMS), according to Article 17 [69]. The GMS must be documented in an

organized and structured way. This should include aspects such as development, testing and post-market monitoring activities.

Article 20 states that providers must take the necessary actions when their systems are not compliant to the AIA or when a risk is presenting. Deployers and distributors of the AI system, authorized representatives and notified bodies, where applicable, must be informed about this by the provider. Furthermore, Article 21 states that providers must fully cooperate with supervisory authorities, including providing all requested documentation. Lastly, providers located outside the EU must appoint an authorized representative to ensure compliance, according to Article 22 [69].

**Conformity assessment**
Article 43 state that providers of high-risk AI systems are obligated to show that they comply with the requirements stated in the AIA. Therefore, the conformity assessment procedure needs to be done. Providers of AI-driven risk assessment models used by HLIC should perform this procedure internally. The procedure consists of the verification by the provider of several articles of the AIA, which are Article 9 until 15, and 17. Furthermore, the provider must verify that the creation process of the system and the post-market monitoring are aligned with the technical documentation. This is referred to in Article 72 of the AIA. When the conformity assessment is successfully conducted, the provider is allowed to write an EU declaration of conformity, which is the required CE marking to its AI system. [69].

**Post-Market Monitoring**
Article 72 states that providers of high-risk AI systems must implement a post-market monitoring system (PMMS), in a way that is proportionate to the size of the AI system and the risks that it introduces. The PMMS should be based on the post-market monitoring plan, which is part of the technical documentation. The PMMS should be able to document and analyze relevant data, that is provided by the deployer, about the performance of the AI system. This data supports the provider in evaluating compliance to the requirements of the AIA [69].

**Reporting of serious incidents**
Providers of high-risk AI systems must report serious incidents to the market surveillance authority of the country that it occurred in, according to Article 73. A serious incident refers to any event or failure of an AI system that results in one or more of the following outcomes [69]:

- The death of an individual or significant damage to the health of an individual.
- A major and irreversible disruption to the functioning or management of critical infrastructure.
- A breach of provisions under EU law that are designed to safeguard fundamental rights.
- Serious damage to property or the environment.

**Digital Omnibus**

On the 19th of November 2025, The European Commission introduced the Digital Omnibus on AI Proposal [17]. This is part of a broader Digital Package, along with the Data Union Strategy and European Business Wallets [4]. This proposal contains amendments to the AIA and is still being negotiated under the EU's legislative procedure by the European Parliament and the Council [23]. Therefore, it is necessary to emphasize that these adjustments are not implemented yet. The European Commission wanted to make these adjustments due to several reasons. The first one is due to a report published by Mario Draghi, former President of the European Central Bank, in 2024. This report was about European competitiveness and included the main message that Europe is falling behind on technical level compared to the U.S. and China. The AIA was the reason why investments in AI development in Europa decreased. Secondly, the assigning of national competent authorities was not fast enough. The writing of the harmonized standards by the European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CENELEC) was delayed, which is further described later this section. For this reason, it will be complicated for providers to check whether they are compliant to the high-risk provisions. The last reason was that many tech companies lobbied the EU regarding the AIA adjustments. Even the Trump administration became involved in pressuring the EU to soften the AIA.

If the proposal would be accepted by the European Parliament and the Council, the major adjustment would be the deferred starting dates of the high-risk AI systems provisions. These starting dates will take effect a specified number of months after the publication of harmonized standards, specifications, and guidelines by CEN and CENELEC for providers of high-risk AI systems. For AI systems listed in Annex III, which include risk assessment models used in life and health insurance, the provisions will apply six months after CEN and CENELEC publish their harmonized standards [4].

#### AI Act criticisms

Many researchers have analyzed the AIA and provided critical reflections on it. First, researchers state that the AIA is very complex and hard to interpret. The articles of this law contain technical and legal jargon, which is hard to understand, especially for deployers with limited knowledge of AI. Furthermore, an implementation gap exists within the AIA, as translating legal requirements into practical applications remains challenging. Articles of the AIA often contain terms such as 'acceptable' and 'appropriate', which are adjectives that can be interpreted differently. This makes it more difficult for IT auditors to determine compliance with the AIA obligations.

Moreover, the many provisions of the AIA introduces two risks, which are a decrease in AI-related innovation and an increase in costs for organizations to ensure compliance. Lastly, researchers argue that providers of high-risk AI systems are highly dependent on harmonized standards that will be published by the CEN and CENELEC, which is further described later this section. These standards provide guidance in complying to the provisions of the AIA [15, 16].

### 4.1.2 GDPR

The GDPR is applied since 2018 and it is one of the most strict international law regarding data and privacy [83]. This EU law was introduced due to the uprising amount of personal data that organizations started to collect and use. Therefore, the GDPR served as a foundation for the AIA. The GDPR contains multiple rules for organizations on how to handle data of citizens [30]. When they do not comply to these obligations, they will receive strict penalties, such as fines reaching to tens of millions of euros. The GDPR states what personal data and data processing include and it defines 3 different roles. This is stated as follows [66]:

- **Personal data**. "Any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person." [66].
- **Data processing**. "Any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction." [66].
- **Data controller**. "the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law." [66].
- **Data processor**. "A natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller." [66].

There are 7 principles that form the foundation of the GDPR. The rules that are stated in the GDPR are based on these values. These principles are as follows [2, 1]:

- **Lawfulness, Fairness, and Transparency**. Firstly, data processing must be lawful. Aspects like consent from the data subject and protection of vital interests are important here. Furthermore, data use may not lead to unjust or detrimental outcomes to individual. Lastly, organizations must be open to individuals when their data is used, including what is used, how it is used and with whom it is shared. They should also know what their rights are.

- **Purpose limitation**. This principle requires organizations to collect personal data only for clearly defined, specific, and legitimate reasons. organizations cannot use it later in ways that do not align with those original purposes. It ensures that individuals know upfront how their data will be used and prevents organizations from applying the data to undisclosed or unintended activities.
- **Data Minimization**. The third principle states that organizations should only collect and process personal data that is truly necessary, ensuring it is appropriate and relevant for the intended purpose. In other words, they must limit data collection to what is essential and avoid gathering any information they do not actually need.
- **Accuracy**. This principle requires organizations to ensure that personal data is correct and, when needed, updated regularly. Incorrect or outdated information can seriously harm individuals, as decisions based on false data may lead to unfair or damaging outcomes.
- **Storage Limitation**. Personal data should only be kept for as long as it is needed to fulfill the purpose for which it was collected. Once the data is no longer required, it must be deleted or anonymized so that the individual can no longer be identified.
- **Integrity and Confidentiality**. This principle focuses on keeping personal data secure. It requires organizations to handle data in a way that protects it from unauthorized access, unlawful use, and accidental loss, destruction or damage. This is done by implementing proper technical and organizational safeguards.
- **Accountability** The accountability principle requires organizations to take ownership of their data processing practices. organizations should be able to prove that they comply with all the other GDPR principles. This can be done by maintaining detailed records of data processing practices.

### 4.1.3 DMA

Since 2023, the Digital Markets Act (DMA) is applied within the EU. This law is meant to ensure fair competition in digital markets by regulating 'Gatekeepers'. Gatekeepers are large digital platforms that provide services in the EU, such as online search engines, app stores, and messenger services. They have to comply with specific obligations and prohibitions imposed by the EU, to create a digital market where smaller digital companies can enter and compete fairly. Article 15 of the DMA states that companies must write a detailed explanation of their profiling techniques, which are often based on AI techniques. Furthermore, this explanation must be audited by an independent auditor and sent to the European Commission. The European Commission can set a methodology or procedure for this audit [67].

### 4.1.4 DSA

The Digital Service Act (DSA) is applied within the EU since 2023 and complements the DMA. The DSA introduces obligations for Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs). The EU aims to create a safe digital environment with this law, in which the fundamental rights of citizens are protected. Article 37 of the DSA states that VLOPs and VLOSEs must undergo an independent audit at least once a year. The auditor must assess compliance with the DSA obligations. The VLOP or VLOSE must provide the auditor with access to all relevant information, including data and systems, and fully cooperate with the audit [68].

## 4.2 Harmonized standards

This section elaborates on the standards that have been, or will be, published by standardization bodies. Regulators encourages organizations to follow these standards, because this leads to compliance to the law. One advantage of the standards is that they provide a consistent way for organizations to comply with the law.

### 4.2.1 CEN/CENELEC JTC21

The EU uses a system where it provides technical standards that act like official guides, helping companies understand how to meet the requirements of EU laws [65]. The European Standards Organizations (ESOs) are responsible for the setup of these standards. The ESOs consist of three independent organizations and can develop standards on their own initiative. Once these standards are formally approved

by the EU Commission, they serve as evidence that a product or system meets the law. This means that if a company follows one of these harmonized standards, regulators and courts will generally assume the company is complying with the related legal requirements. Therefore, these harmonized standards reduce the EU's compliance monitoring efforts, as organizations apply consistent terminology and procedures.

For the development of the AIA standards, two organizations are appointed to work together on this. These are the European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CENELEC), which are combined to form the committee CEN and CENELEC Joint Technical Committee 21 (CEN/CENELEC JTC21) [11].

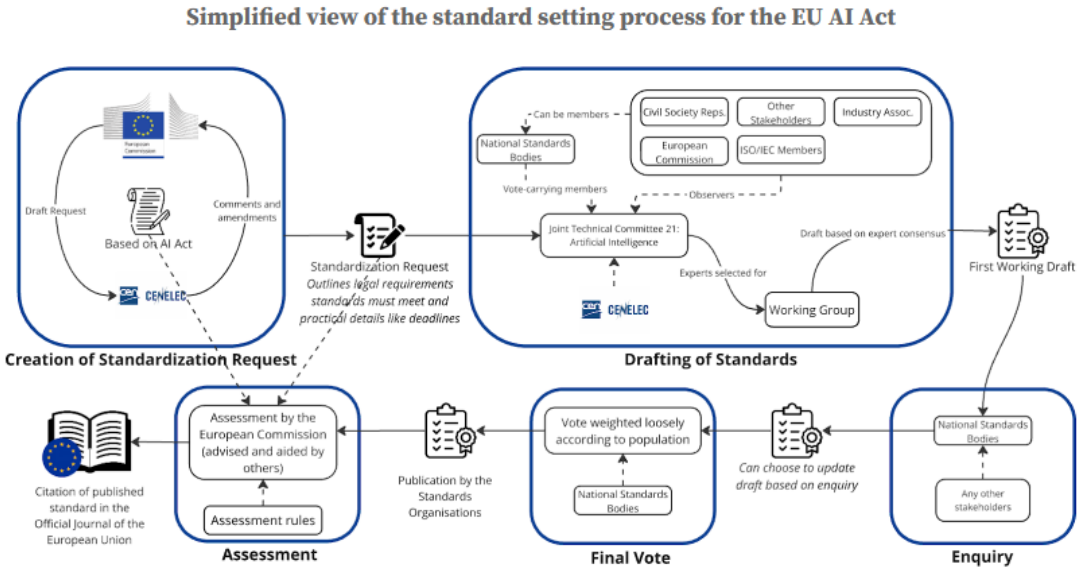



Figure 4.1: A visualization of the development process of EU standards [65].

The working process of developing technical standards consists of five phases. First, the Commission gives a request to the ESOs to create a standard. When this request is accepted, the ESO will provide a draft of the standard. This will be done by technical experts from National Standards Bodies (NSBs) who do this work voluntarily. Usually, these standards align with the standards that are written by the ISO/IEC. During the enquiry stage, national standards bodies review and vote on a draft standard, providing comments that the working group must address. In the next step, the NSBs give their final vote about the updated version of the standard, to either accept or reject it. If the standard is accepted, it will be published by CEN and CENELEC. Finally, the standard is assessed by the EU Commission. The standard will be published in the Official Journal of the European Union (OJEU) if this assessment was successful [65]. This whole process is visualized in figure 4.1.

At the time this article was written, 21th of July 2025, the CEN/CENELEC JTC21 was still busy with the draft for the AIA harmonized standards [65]. These harmonized standards contain guidelines for compliance, and examples of use cases of high-risk and not-high-risk AI systems [69]. Although the EU Commission requested the AIA harmonized standards in 2021 and CEN/CENELEC JTC21 had the first meeting in June 2021 about this, the work is still not done. According to news reports, the harmonized standards will probably be finalized in 2026 [21].

### 4.2.2 ISO/IEC

The International Organization for Standardization (ISO) [36] and the International Electrotechnical Commission (IEC) [34] are two important international organizations that publish standards that are commonly used by organizations worldwide. These standards are made by different experts that work together to provide these. The ISO focuses on many industries, with its primary goal to ensure quality, safety and efficiency. The IEC focuses more on (electro)technical-related issues. New technologies require new guidelines regarding the use of these, to ensure safety for example. In 2023 [36], the ISO/IEC published the ISO/IEC 42001:2023 together. This is currently one of the most used standard for the use of an Artificial Intelligence Management System (AIMS). The ISO/IEC 42001:2023 provides guidance

to organizations that develop AI systems or use AI systems in different kinds of workflows, so they can maintain their systems appropriately. Adopting this standard ensures that AI is used in a safe, ethical and transparent manner. The objectives of the ISO/IEC 42001:2023 are based on the values of RAI. For this reason, this standard is frequently used by auditors when auditing AI systems [9]. At the same time, organizations can receive a certificate for completing an audit on the ISO/IEC 42001:2023.

Furthermore, the ISO/IEC introduced other standards regarding the use of AI, which are described as follows [36]:

- **ISO/IEC 22989**. This standard provides terminology and concepts in the field of AI. This shared vocabulary can be useful for organizations that use AI but are not very known with the commonly used terminology.
- **ISO/IEC 23894:2023**. This standard offers practical guidance for organizations that develop or deploy AI systems, on managing AI-related risks [39].
- **ISO/IEC 38507:2022**. This standard sets out recommendations for the directing roles within organizations, helping them oversee and steer the use of AI [38]. Its purpose is to support organizations in using AI in a way that is effective, efficient, and aligned with acceptable practices.
- **ISO/IEC 42005:2025**. This standard offers support for organizations to perform an impact assessment on their AI systems [40]. This process helps organizations to see how outcomes of the AI system could impact individuals, communities, or societies.

### 4.2.3 IEEE

IEEE is one of the largest technical professional organization [35]. It is mostly known for the published technical standards. However, the IEEE is also a leading scientific journal and paper publisher. The goal of this global community is to share knowledge about advanced technologies for the benefit of humanity. The IEEE has also published many standards regarding the use of AI. These standards belong to the IEEE 7000 series, which is described in this section.

The IEEE 7000 series provide a collection of standards that address ethical and societal challenges within the use of autonomous and intelligent systems [5]. One of the standards of the IEEE 7000 series is the IEEE P7001, for example [53]. This standard focuses on improving transparency in autonomous technologies, such as self-driving cars and drones. The AI systems that are embedded in these, cause a black box model problem when choices are made by them. This unclear decision-making damages the user trust and safety when using these AI-driven machines. Because of this, the IEEE P7001 was introduced. This standard sets up measurable and verifiable transparency criteria, using a multi-tiered approach to match explanations to the needs of different stakeholders. These stakeholders are:

- **Non-expert users**. These could be drivers or passengers for example. They only receive simple explanations.
- **Expert stakeholders**. These could be auditors or regulators for example. They receive full technical documentation, systems logs and decision rationale evidence.

The standard provides 5 levels of transparency explanations [53]. These range from basic explanations to extra extensive explanations. This level is determined depending on the knowledge level of the user or stakeholder. The standard is also known for its scalability and flexibility, so it can evolve with technical developments. This also helps to not overburden simple AI systems, while setting higher requirements to critical ones.

Another relevant standard is the IEEE P7003. This standard is about algorithmic bias considerations. The IEEE p7003 provides protocols to avoid negative bias in algorithms, including guidelines for data selection and testing. Other standards of the IEEE 7000 series contain guidelines for the design, data governance, and ethics regarding autonomous and intelligent technologies.

## 4.3 Frameworks

This section consists of AI frameworks from five organizations. These AI frameworks provide guidance in AI-related topics, which may support compliance with the AIA. These international and professional AI frameworks are non-binding. Organizations can decide for themselves whether to adopt these AI

frameworks. However, adherence to these AI frameworks does not necessarily imply compliance with the AIA. Researchers emphasize that it is necessary to develop different frameworks for different AI systems and different sectors. Many AI systems use different algorithm techniques, datasets, and processes. Every sector struggles with different sorts of risks, which all have different impacts on society and individuals [53, 27].

### 4.3.1 NIST

The National Institute of Standards and Technology (NIST) is an institute from the U.S. that establishes standards, guidelines and frameworks in many scientific fields [59]. The most relevant framework that NIST developed regarding AI is the AI Risk Management Framework (AI RMF) [58]. This framework argues that AI risk management is a crucial element to provide RAI. Key aspects for RAI are sustainability, social responsibility, and human centricity. The framework is organized into two main sections. The first section explains how organizations can conceptualize and structure the risks associated with AI, and it clarifies who the framework is intended for. The AI RMF is meant for organizations to help them identifying, evaluating, and mitigating risks that arise during the design, development, and deployment of AI. It then examines AI risks and trustworthiness, and explains the qualities that define responsible AI systems. These qualities are listed as follows:

- Validity and reliability
- Safety
- Security and resilience
- Accountability and transparency
- Explainability and interoperability
- Privacy
- Fairness

Part 2 contains the core of the framework. This part outlines four key functions designed to help organizations manage AI-related risks in practical settings. These four functions are Govern, Map, Measure, and Manage. The Govern function includes all phases of an organization's AI risk management activities and overarching procedures. The functions Map, Measure, and Manage are applied within the context of specific AI systems and at particular points in the AI lifecycle [58].

Furthermore, NIST published the NISTIR 8312, which is a report that outlines four key principles for explainable AI [73]. These principles could guide organizations that develop or deploy AI systems in ensuring AI explanations to be trustworthy, meaningful and aligned with user needs. The four principles are outlined as follows:

- **Explanation**. This means that the AI system provides related proof or reasons for its outcomes.
- **Meaningful**. This means that the AI system's explanations are understandable for the intended user.
- **Explanation Accuracy**. This mean that the level of accuracy of the AI system's explanations should be significant.
- **Knowledge Limits**. This means that the AI system should run only in predefined conditions and only when it is confident enough in the accuracy of its outcomes.

The report recognizes the fact that every AI system needs a different way of explaining. One of the reasons for this is that the intended user can differ. For instance, developers and designers of AI systems need other explanations than policy makers and end users. Another reason is that the different fields in which AI is used, require different ways of explaining. AI systems that are used for credit scoring or insurance pricing require a different explanation than AI systems used for entertainment. This is due to the fact that the impact on the user differs [73].

### 4.3.2 NOREA

NOREA is the Dutch professional body for IT auditors [60]. This organization keeps an official register of qualified IT auditors (REs) so that clients and other parties can easily verify whether someone meets the professional standards based on their education and experience. The goals of this organization is to represent the collective interests of the members, and to improve the quality of professional practice and development of the field. NOREA published a guideline that includes principles for responsible AI investigations [20]. These principles were created support Dutch registered IT auditors in performing audits of AI systems used within companies, based on the key features of responsible AI. The principles consist of 119 considerations, which are categorized into five risks and the six stages of the Cross-Industry Standard Process for Data Mining (CRISP-DM), with Governance added as an additional category. The risk categories are Governance, Ethics, Privacy, Performance, Security, and Roles & Responsibilities. The CRISP-DM stages are Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Each consideration consists of at least one key question that can be asked about the AI system.

### 4.3.3 KPMG

The third-party audit firm KPMG has published several frameworks and guidelines for auditing and assessing AI systems [46]. Some of these are publicly available, and some of these are confidential and only meant for KPMG employees.

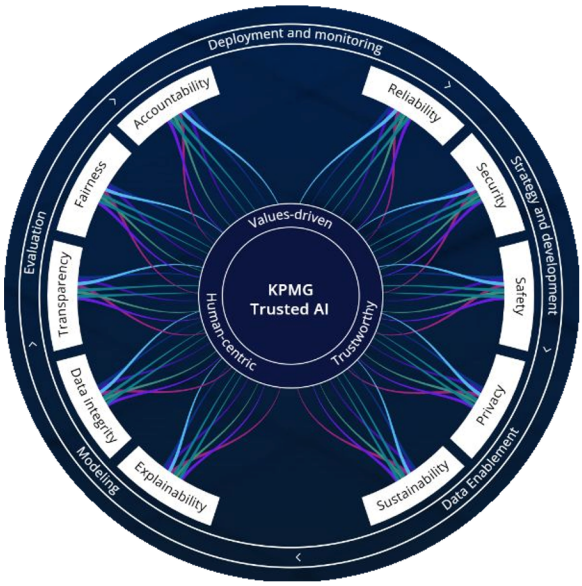



Figure 4.2: The KPMG Trusted AI framework [42].

One the publicly available frameworks is the KPMG Trusted AI framework [42]. This framework describes the values that KPMG finds crucial when designing and developing AI systems. KPMG sees AI as a powerful tool for progress and works to combine technological innovation with human expertise to create trustworthy, safe, and fair solutions. This approach aligns closely with the principles of RAI. This commitment to RAI drives the firm to openly share its guiding principles, ethical foundations, and governance practices. The values that KPMG finds crucial in designing an developing AI are illustrated in figure 4.2. The KPMG Trusted AI framework consists ten values, which can be described as follows [42]:

- **Reliability**. AI systems must always perform according to their defined purpose and boundaries, and deliver the expected level of accuracy.
- **Security**. Technical measures that improve robustness and resilience should be implemented to protect AI systems from malicious actors, false information, and unexpected disruptive events.
- **Safety**. AI systems should be created and deployed in a way that prevents harm to individuals and organizations.

- **Privacy**. AI systems should be built to meet all applicable privacy and data protection regulations.
- **Sustainability**. AI systems should be developed in a way that prioritizes energy efficiency, lowers carbon emissions, and contributes to a more sustainable environment.
- **Explainability**. AI systems should be able explain clearly how a conclusion was made and why that outcome makes sense.
- **Data integrity**. Data used in AI systems should be collected in line with all relevant regulations, and evaluated for precision, suitability, and quality to ensure trustworthy outcomes.
- **Transparency**. AI systems should include transparent reporting practices that give stakeholders clear insights of the functionality of the system.
- **Fairness**. AI systems should be developed to minimize or remove bias toward natural persons and communities.
- **Accountability**. Throughout the entire AI lifespan, human oversight and responsibility must remain actively involved so that risks are properly controlled and legal requirements are consistently met.

Another framework explains what value internal audits could add to the external audits of AI systems [43]. Moreover, this framework provides market insights from KPMG teams. This part states that many audits will focus on AI governance and AI risk management, and many auditors feel like their not prepared enough to perform these audits. Lastly, the framework provides examples of AI audits and assessments that have been conducted by KPMG globally. These examples are as follows:

- **AI governance model**. Assessing the general AI governance and risk management structure that leads AI projects within an organization.
- **AI framework and policy**. Assessing the processes, frameworks and policies guiding the development and use of AI within an organization.
- **AI governance workshops**. Interactive sessions with relevant stakeholders to teach them AI governance practices.
- **Data governance reviews**. Assessing data governance and management processes related to AI and examine compliance with regulations regarding data governance.
- **AI solution pre/post deployment assessment**. Evaluating the AI's systems maturity, functionality, and influence on operations before and after deployment.
- **Third-party/Vendor management reviews**. Assessing the procedures for selecting an managing AI providers.
- **AI security and privacy**. Assessing security and privacy risks of AI systems and ensuring these are mitigated as much as possible.
- **AI use case intake process**. Assessing the procedures for defining, prioritizing, approving AI use cases within an organization.

### 4.3.4 IIA

The Institute of Internal Auditors (IIA) has introduced a framework for the auditing of AI systems [6]. This framework guides internal auditors who need to provide advisory and assurance services regarding AI systems. The framework consists of AI knowledge and a checklist with requirements for AI assessments. It is frequently updated so it can keep up with the continuous developments in the field of AI. The framework emphasizes the importance of the four characteristics of an AI system: transparency, explainability, responsibility, and auditability.

### 4.3.5 ForHumanity

ForHumanity is a non-profit public organization that provides guidance in identifying, assessing and mitigating risks in the field of AI and autonomous systems. These are risks regarding ethics, bias, privacy, trust, and cybersecurity. ForHumanity is mostly focused on the development of Independent Audit of AI Systems (IAAIS) [25]. The IAAIS consists of developed audit criteria, rules, and certification schemes so

that AI systems can be audited independently, transparently, and consistently across jurisdictions and sectors. The 5 IAAIS Audit Rules are described as follows:

- **Binary - compliant/non-compliant**. A control must result in a compliant or non-compliant conclusion. Otherwise, a gray area arises where auditors may interpret controls differently, which increases liability risks. Binary controls can be turned into automated checks and integrated into compliance-by-design systems. This also creates clarity and trust for stakeholders.
- **Measurable and unambiguous**. Every control must be objectively measurable and clearly defined. This makes auditing consistent and repeatable. Furthermore, case studies and examples should be applied so that controls are better understood. Forums must exist to help each other in understanding the control.
- **Iterated and open-source**. Every control is reviewed multiple times and every thought and idea about a control will be included in the audit process. The quality and fairness of the audit will be ensured because of this. This is a transparent process that is repeatedly performed.
- **Consensus-driven**. No control can be universally accepted. Therefore, people should create a consensus, which balances practicality with inclusiveness. This can be done by holding discussions that help mediate dissent.
- **Implementable**. Every control must be implementable. This means that controls should be realistic and easy to implement into a system. Auditors should rather audit a small and practical control than a large and unrealistic one.

ForHumanity also published a certification scheme to guide providers of AI systems in auditing under the AIA [26]. Besides the AIA, the audit criteria in this scheme are aligned with regulations such as the GDPR, DSA and Accessibility Act. The certification scheme is eligible for both high-risk and limited risk AI systems. The framework consists of criteria covering most of the requirements of the AIA.

# Chapter 5

# Requirements and specifications

In this chapter, the requirements and specifications for the AI audit framework are stated and explained. The requirements and specifications are derived from both the literature review and the analysis of the first interviews. These exploratory interviews were analyzed with the open coding and thematic coding method. This is elaborated in the next section.

## 5.1 Literature review

In Chapter 4, we observed that the AIA sets many obligations for AI-driven risk assessment models used by HLIC. The obligations include areas such as risk management, data governance, technical documentation, record-keeping, transparency, human oversight, accuracy, cybersecurity, monitoring, and reporting. Therefore, it is necessary that the AI audit framework include controls that cover all these previously mentioned obligations. At the same time, the AIA requirements are abstract, which makes them difficult to operationalize in practice. Therefore, an important specification should be to implement measurable controls within the AI audit framework.

Moreover, existing standards and frameworks from ISO/IEC, IEEE, NIST, NOREA, KPMG, IIA, and ForHumanity provide useful guidance on RAI-related principles. These are principles such as transparency, data governance, fairness, security, accountability, and accuracy, which can also be considered as important AIA principles. These principles are therefore relevant for the development of our AI audit framework. In addition, frameworks of NIST, NOREA, KPMG, IIA, and ForHumanity include methods and criteria for AI auditing. These elements can also be used for the development of our AI audit framework. On the other hand, we observed that none of the existing standards or frameworks specifically addresses the auditing of high-risk AI systems under the AIA. Moreover, no standard or framework exist for auditing AI-driven risk assessment models used in health and life insurance. These two observations emphasize the relevance of our AI audit framework.

## 5.2 Exploratory interviews

The exploratory interviews were conducted as described in Section 3.1.3. The interview answers were first analyzed with the open coding method within Excel. All the answers from the interviews were categorized into 15 codes. We defined these 15 codes after reading all the interview answers in advance. The codes were based both on the topics of the interview questions and on additional topics that the interviewees talked about. After we assigned every sentence to a code, we used the thematic coding method to analyze the answers. For this, 11 themes were chosen from the 15 previously derived codes. We chose these themes based on the relevance to the research and the artifact, and because these themes corresponded with a relatively large number of interview responses. For every interviewee and every theme, we assigned a summarized insight to the corresponding cell in the data matrix. This data matrix is attached to Appendix A.3. Table 5.1 contains each theme that we chose for the thematic coding. For every theme, an explanation is provided why we chose this theme.

| Theme | Explanation |
|---|---|
| AI audits done | This refers to everything that is related to AI audits that are already done by the ITAA department of KPMG. This consists of different engagements, such as audit and assurance engagements, and supporting internal projects of other organizations. This is important, as it provides answers to the first sub-question. Furthermore, it explains the relevance of this research. |
| AI audit process | This refers to how the engagements regarding AI is performed. Information about what is audited or assessed exactly, is included here. This provides information about the process of auditing or assessing AI, which helps in answering sub-question 1. |
| Challenges | This refers to everything that is related to the challenges that arise due to AI auditing. This consists of gaps and other difficulties regarding AI auditing within ITAA and outside KPMG. This is relevant for our AI audit framework, as it is necessary to know what challenges occurred, to determine the requirements and specifications for the framework, which answers sub-question 2. |
| Frameworks existence | This refers to everything that is related to the frameworks, standards, and guidelines that already exist for AI auditing. These can be published by private or public sector organizations, or Non-Governmental Organizations (NGOs). This is relevant for the construction of our framework, as the framework could be build upon other frameworks that already exist. Therefore, this information helps answering sub-question 1 and 2. |
| Frameworks characteristics | This refers to the characteristics of the frameworks, standards, and guidelines developed for AI auditing. This include the capabilities and limitations of the frameworks, standards, and guidelines. This is relevant, as it may contain information that can be translated to requirements and specifications for our audit framework, which answers sub-question 2. |
| Auditability | This refers to everything that is related to the auditability of AI systems. This consists of the current experiences of auditors when auditing AI systems. This theme is important, as it gives us knowledge about how well auditors could audit the AI systems of their clients. The process of providing evidence for auditing, is also necessary for our audit framework. Therefore, this helps answering sub-question 2. |
| AI skills | This refers to everything that is related to the current AI knowledge and skills of the auditors of ITAA, and how the interviewees look at this situation. This is relevant, as it describes how well prepared the employees are for AI auditing. There could be an audit framework to use, however, this cannot be properly used if the employee does not know anything about AI. Therefore, this information helps answering sub-question 2. |
| HLIC audit | This refers to everything that is relevant when auditing AI-driven risk assessment models used by HLIC under the AIA. This is important for our audit framework, since we want to be able to audit these sort of models with this framework. Therefore, the answers assigned to this theme could be translated to requirements and specifications for our audit framework, which answers sub-question 2. |
| Demand | This refers to everything that is related to the demand of AI audits under the AIA in the future. There is uncertainty about what the future of AI auditing under the AIA will look like. Due to the experiences of the interviewees, we could gain insights about this. This may highlight the need for an AI audit framework. |
| Risks | This refers to everything that is related to the risks that may be introduced if AI auditing is not properly executed within our society. This consists of risks for companies, as well as for our society. Similarly, this could underscore the need for an AI audit framework. |
| Future changes | This refers to everything that is related to the necessary changes for the future. This consists of regulatory and organizational changes, both within KPMG as outside KPMG. This is relevant, as it could provide requirements or specifications for our audit framework, which helps answering sub-question 2. |

Table 5.1: This table describes every chosen theme for the thematic open coding and explains why this theme is relevant.

## 5.3 Contribution arguments

As earlier described in Section 3.1.2, exploratory interviews were conducted to obtain instrument design goals. Besides this, a literature review is done to determine the instrument design goals. Wieringa et al. [82] emphasizes that within a design problem, similar to this research, there is no best solution. For every stakeholder, this depends on whether its goals are translated into requirements for the artifact. Stakeholders typically do not specify the requirements directly. In this problem context, the stakeholders are the auditors, the clients, and the regulators. Translating the stakeholder goals into requirements is done by the design researcher. These requirements can be justified by contribution arguments. A contribution argument (CA) is a reason why the framework that meets a certain requirement contributes to a resolving a problem within a context. We chose to use contribution arguments for our research, as we can use this systematic method for translating the stakeholders needs into actionable framework requirements and specifications. In eq. (5.1), we present the formula of a contribution argument [82].

$$\text{(Artifact requirement)} \times \text{(Context assumption) contribute to (Stakeholder goal).} \tag{5.1}$$

There is a possibility that multiple contribution arguments exist for one requirement. Furthermore, a contribution argument is a prediction, as it suggests that when the artifact is applied within its problem context, the interaction will support the achievement of stakeholder goals. However, this is based on a context assumption that cannot be stated with certainty, which makes it fallible. From the exploratory interviews and the literature review, the contribution arguments for the AI audit framework are derived. In table 5.2, the contribution arguments are stated per row.

| ID CA | Artifact requirement | Context assumption | Stakeholder goal |
|---|---|---|---|
| CA1 | The framework contains all the controls relevant for auditing providers of high-risk AI systems. | Providers must comply with all obligations associated with their risk category. | Every provider of high-risk AI systems can be audited for compliance with the relevant obligations. |
| CA2 | The framework follows standardized structures and practices. | Demand for AIA audits will increase. | IT auditors can perform AIA audits in a standardized way effectively. |
| CA3 | The framework supports sector-specific customization. | AI systems differ significantly across sectors. | Different sorts of AI systems can be audited under the AIA. |
| CA4 | The framework includes measurable controls. | AI systems must be audited on multi-interpretable obligations. | IT auditors can audit every AIA obligation with measurable controls. |
| CA5 | The framework provides structured guidance, including examples, for IT auditors with limited AI expertise. | General IT auditors will have to audit AI systems. | Every IT auditor is able to perform an AIA audit effectively. |
| CA6 | The framework is able to produce a risk classification of the AI system. | The risk classification stays the key indicator for the AIA obligations. | IT auditor can check whether the correct risk assessment is done by the client and correct obligations are applied. |
| CA7 | The framework must be able to audit fairness, bias, and data drift within risk assessment models used by HLIC correctly. | The market surveillance authority considers fairness violations to be the most crucial aspect to monitor. | IT auditors can audit AI systems on its fairness, bias, and data drift. |

Table 5.2: This table contains which artifact requirement and context assumption contribute together to the stakeholder goal.

# Chapter 6

# Synthetic data

In this chapter, we present the construction of the synthetic data that forms the mock demonstrator case of this research. Then, we describe the fictive case, including the evidence that is provided for the mock demonstrator case. This is used to evaluate the designed AI audit framework.

## 6.1 Case construction

This synthetic data is not based on real client data, due to confidentiality obligations. However, the construction of the case study has a expert-driven design. The synthetic data is inspired by real IT audits that have been conducted by KPMG. KPMG has conducted many audits for insurance companies, which is also a reason why we chose a model that is used in this sector in the first place. We used prior audit methods and reports, the procedures for gathering evidence, and the AIA requirements for high-risk AI systems to create the synthetic data. Therefore, the synthetic data resembles real cases and is suitable for an AI audit under the AIA. For the visualizations of the synthetic data, Copilot and Canva was mainly used. These visualizations are part of the evidence that NexaSure provides to the auditors for the mock demonstrator case, which consists of an AI audit with the upgraded AI audit framework. After the synthetic data was constructed, it was validated by an IT auditor on its realism. This validation included criteria such as completeness, clarity, and alignment with real cases. The IT auditor stated that these criteria were sufficiently met.

### 6.1.1 Case: NexaSure

The case study is about KPMG performing an AI audit for the fictive firm NexaSure. NexaSure is an AI development company that is founded in 2009 and which headquarters are positioned in Leiden, the Netherlands. The company consists of 1200 employees and the annual revenue lies around 20 million euros. The firm is primarily focused on developing AI-driven risk assessment models for clients in regulated industries, such as insurers, pension funds, and banks.

#### Risk assessment model

The recently developed risk assessment model of NexaSure, that can be used by insurers for underwriting, is referred to as RiskQuant v3.2. RiskQuant v3.2 is able to predict the risk of a new contract with an applicant. In other words, it can predict if an applicant will file relatively many claims or not. This model aims to support the underwriting processes of health and life insurance companies, provided that such underwriting is legally permitted in the relevant country. However, before RiskQuant v3.2 can be placed on the market, it needs to be compliant with the AIA. RiskQuant v3.2 uses gradient boosted trees to predict the risks of accepting new applicants and are trained with tabular data. Rule-based constraints are used to make sure the model never produces outputs that violate established actuarial logic or regulatory requirements.

NexaSure uses different data sources for the training and testing of RiskQuant v3.2. The first one is data provided by the applicant itself. Besides personal data, this data consists of prior medical claims, and whether the applicant smokes. The outputs of the model consist of primary output and secondary

outputs. The primary output is the risk score that is assigned to an applicant. This score is a number between 0 and 100, and is used to determine the decision class:

- **0-40**. Underwriter sets a low premium.
- **41-60**. Underwriter sets a medium premium.
- **61-80**. Underwriter sets a high premium.
- **81-100**. A mandatory human review has to be performed. With this, an underwriter must manually assess the applicant before any decision is made. This decision is either to decline the applicant, set a high premium, or set a lower premium than the model recommended.

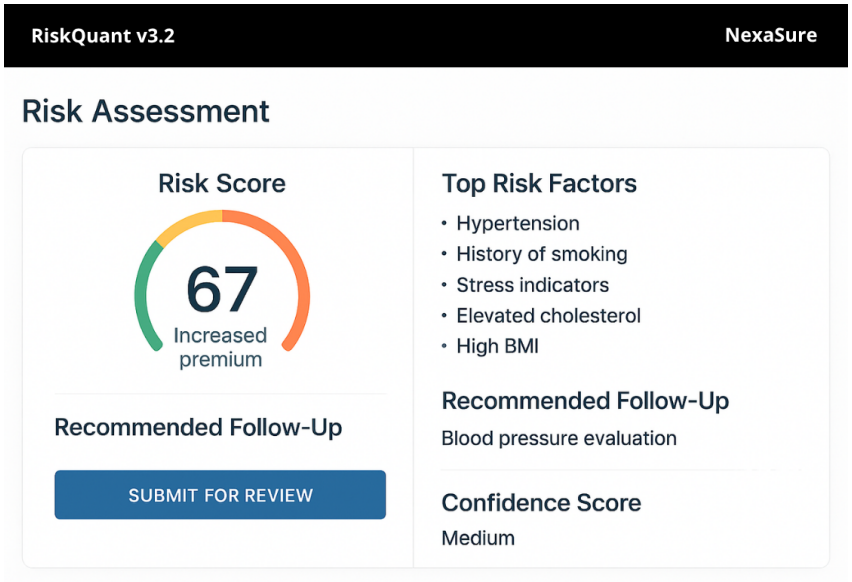



Figure 6.1: User-interface of the RiskQuant v3.2 model when the output is given.

The secondary output consist of supporting information to understand why the primary output is given by the model. This information contains a top five of the risk drivers that influenced the model's choices. These risk drivers explain why the model gave certain risk scores. Furthermore, recommended medical follow-up checks are provided by the model. This could include requests for a blood test or physician's statement, for instance. This helps underwriters focus on cases where more information is needed. Lastly, confidence scores are given, which indicates how certain the model is about its output. This helps the underwriter to judge the reliability of the model. In figure 6.1, a screenshot is shown to illustrate what the model produces as output, when predicting the risk score of an applicant.

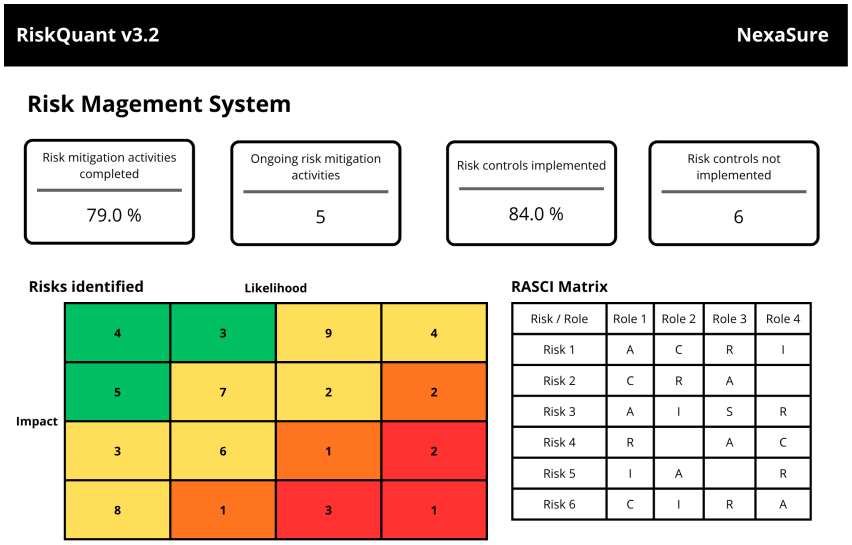



Figure 6.2: The user-interface of the RMS of NexaSure.

Furthermore, NexaSure has implemented an RMS, from which the user-interface is shown in figure 6.2. This image shows that NexaSure has a clear overview of its risks, risk mitigation activities and controls, regarding the risks that arise due to the use of RiskQuant v3.2. Besides this, NexaSure has an overview of the roles and responsibilities within the company, regarding these risks. This overview is illustrated within the RASCI Matrix.

Moreover, NexaSure provides a technical documentation file that consists of all the needed information to meet the AIA requirements regarding technical documentation. The technical documentation of NexaSure is attached to Appendix B.

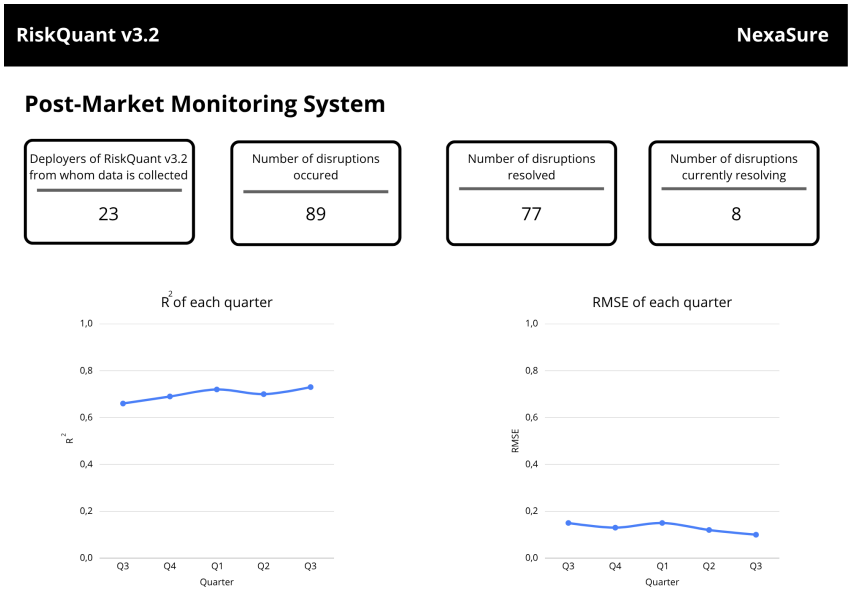



Figure 6.3: The user-interface of the Post-Market Monitoring System (PPMS).

Finally, the Post-Market Monitoring System (PMMS) is implemented by NexaSure. This system is able to monitor the functionality of the system and is proportionate to the risks regarding health and life insurance. Every malfunction that occurred or risk that has been realized is registered here, which helps Nexasure manage and resolve these disruptions. Deployers of the risk assessment model provide system

data to the PMMS, which supports NexaSure in being compliant with the AIA continuously. This data includes the $R^2$ and RMSE score of the risk assessment model. This is measured per quarter, since the model is retrained every quarter. The PMMS is illustrated in figure 6.3.

# Chapter 7

# Design

In this chapter, we present the design activity of the AI audit framework. The purpose of the audit framework is to audit AI systems that fall under the high-risk category of the AIA. The focus lies on high-risk AI systems due to its huge amount of obligations compared to the limited risk category, which makes the audits of high-risk AI systems most relevant. Moreover, the framework is developed for auditing AI systems, even though this has not commonly been done before. During the interviews, experts stated that they have primarily assessed AI systems rather than audited them, as there has been no obligation for organizations to audit their AI system under the AIA. However, researchers and IT auditors argue that this obligation is necessary to create a safe and efficient AI auditing ecosystem. To the interview question "What additional changes are needed to stimulate a well-functioning AI auditing ecosystem? For example, are organizational or legal adjustments required?", expert 1 answered:

> *"What mainly matters is that the AI Act establishes an obligation to have AI systems audited."*

Expert 4 answered the same question as follows:

> *"I wonder whether, in the future, an audit or regular checks of your IT systems will become just as important as an audit of your financial statements. This is also because we are seeing risks in our society increasingly shift toward technological risks. So perhaps making something like this mandatory could be the solution."*

Due to significant pressure from various stakeholders to introduce an obligation to audit AI systems under the AIA, and the lack of shared proper frameworks for AIA audits, we chose to develop a framework focused on auditing rather than assessment.

The choices made for the design activities of the framework are justified based on the artifact requirements of the seven CAs. To begin with, the AI audit framework is developed in Excel. This is because many IT auditors perform their daily work in Excel, making them highly proficient in using this tool. Furthermore, we chose to develop a control table in Excel, drawing on concepts from the course *Managing Software Evolution*, taught by Prof. Dr. Ir. Joost Visser. The design of the framework consists of two components, which are the preparation process and the audit process. By implementing the framework in this way, it meets the artifact requirement of CA2, as it follows the same standardized structure as a typical IT audit.

The AI audit framework design is based on the four elements discussed by Lam et al. [48], as were described in Section 2.2. We aimed to implement a standardized set of evaluation criteria, with the purpose to check compliance with the AIA. In addition, we consider AI trainings to be a crucial element of the framework. Finally, we emphasize the importance of making the audit report publicly available, as this stimulates transparency and accountability.

## 7.1 Preparation process

The first component, which is the Preparation process, consists of a structured checklist with five questions. Before the audit can be conducted, the IT auditor that leads the audit needs to answer these questions. Most of the time, this 'engagement lead' chooses which IT auditors join the team to help

with this engagement. Therefore, we say that the questions need to be answered by the engagement lead from now on, but this could be different person. These questions relate to the scope of the audit and the requirements that the IT auditor needs to meet. This preparatory phase is essential, as it ensures that the engagement lead develops a clear understanding of the context in which the AI system operates and the objectives of the audit. This information determines which risks of the AI system and which elements of the obligations are relevant for the audit. Furthermore, the preparatory phase provides a clear understanding of which IT auditors are sufficiently well-informed to audit an AI system in a specific sector. These specifications are derived from both the literature review and the exploratory interviews. The literature review shows that every audit is dependent on the specific scope, because the sector, the role of the client, and the intended use of the AI system are the main indicators for the introduced risks and corresponding obligations. Furthermore, academic papers argue that every audit is different due to the different sorts of AI systems that exist, and that organizations often lack AI expertise. During the exploratory interviews, the experts also emphasized the need for determining the scope correctly, as this determines the risk classification and the corresponding obligations. In addition, the experts stated that most of the IT auditors do not possess sufficient AI expertise, although this is necessary to effectively audit AI systems under the AIA. Existing AI audit frameworks and the experts also indicated that IT auditors require knowledge about the sector that the audited AI system is used in.

| AI audit preparation | | |
|---|---|---|
| **Scope** | **Answer** | **Action** |
| 1. What is the field in which the AI system is used? | Essential private services and essential public services and benefits | |
| 2. Does the intended use of the AI system pose a risk to the health, safety, or fundamental rights of individuals? | Yes | The AI system is high-risk and needs to be audited within the context of Essential private services and essential public services and benefits. |
| 3. What is the role of the client? | Provider | The client needs to be audited for compliance with the obligations applicable to a Provider. |
| **Auditor requirements** | | |
| 4. Is the auditor well-informed about current AI theory and practices? | No | The auditor requires training in AI. |
| 5. Is the auditor well-informed about the field that the AI system is used in, including the relevant risks? | No | The auditor requires training in the field of Essential private services and essential public services and benefits. |
| Conclusion scope: | The client needs to be audited for compliance with the high-risk obligations applicable to a Provider, within the context of Essential private services and essential public services and benefits. Only audit elements related to this context should be included in the audit framework. | |
| Conclusion auditor requirements: | The auditor requires training in AI and in the field of Essential private services and essential public services and benefits, including the relevant risks within this field. | |

Figure 7.1: The Preparation process component filled in for a possible NexaSure audit.

The Preparation process component consists of a table with three columns, as shown in figure 7.1. The checklist is filled in for a possible NexaSure audit where the auditor is not well-informed about both the current AI theory and practices, and the field of essential private services and essential public services and benefits. In the left column, the preparation question is asked to the engagement lead. The engagement lead can choose an answer in the drop down menu in the middle column. Except for question one, every answer results in a specific answer that tells the engagement lead what action is needed. This action will automatically appear in the column on the right. These answers are automatically summarized in the 'conclusion scope' and 'conclusion auditor requirements' underneath the table.

### 7.1.1 Scope of the audit

The first three questions aim to determine the scope of the audit. These questions relate to the field where in the system is intended to be used, the risks that are introduced due to its use, and what the

role of the client is. Question 1 is stated as follows: "What is the field in which the AI system is used?". Possible answers to this question are fields that are mentioned in the Annex III, which are listed as:

- Safety components of products
- Biometrics
- Critical infrastructure (physical, digital, or utilities)
- Education and vocational training
- (Self-)employment and workers management
- Essential private services and essential public services and benefits
- Law enforcement
- Migration, asylum and border control management
- Administration of justice and democratic processes
- Other

Furthermore, question 2 follows as: "Does the intended use of the AI system pose a risk to the health, safety, or fundamental rights of individuals?", which can be answered with 'Yes' or 'No'. If the answer to question 1 is not 'Other' and the answer to question 2 is 'Yes', the action that follows is: "AI system is high-risk and needs to be audited within the context of [Answer question 1].", as these are the two conditions of AI systems that fall in this category. If the answer to question 1 is 'Other' or the answer of question 2 is 'No', the following text in the action column will appear: "AI system is not high-risk.". Then, question 3 follows as: "What is the role of the client?". Possible answers to this question are:

- Provider
- Deployer
- Authorized representative
- Importer
- Distributor
- Deployer and Provider
- Deployer and Authorized representative
- Deployer and Importer
- Deployer and Distributor

If the AI system of the client is high-risk, the following action will appear in the table column on the right: "The client needs to be audited for compliance with the obligations applicable to a [Answer question 3]." If the AI system is not high-risk, nothing will appear in the action column.

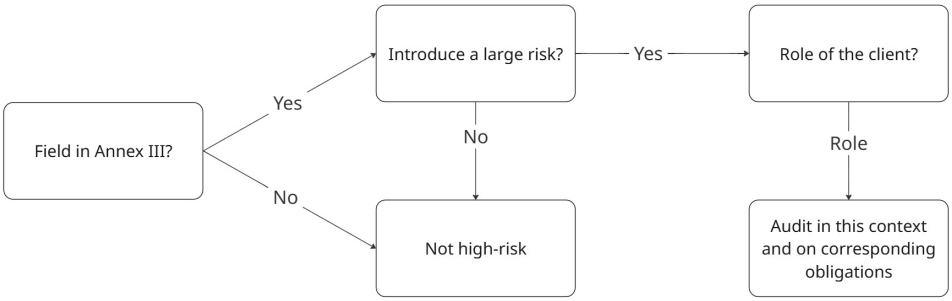



Figure 7.2: A visualization of the audit scope part in the preparation process.

A visualization of the audit scope part is shown in figure 7.2. By determining the scope of the audit within the framework, the framework meets the requirements of CA3 and CA6. The answers to these questions

form a risk classification that determines whether the AI system is high-risk. Together with the role of the client, the scope for the audit is set. These answers should automatically result in a sector-specific customization of the framework, as these are the key indicators that determine the required elements of the audit. This last function is not fully implemented within the framework, as the focus of this research is on the auditing of AI-driven risk assessment models used in HLIC. Therefore, the developed audit controls can only be applied to these AI systems.

### 7.1.2 Auditor requirements

The fourth and fifth questions are about the requirements that the IT auditor needs to meet before auditing the AI system. Question 4 is stated as follows: "Is the auditor well-informed about current AI theory and practices?", which can be answered with 'Yes' or 'No'. If question 4 is answered with 'Yes', the cell in the right column will be filled with: "The auditor does not require training in AI.". If this question is answered with 'No', the following action appears in the column on the right: "The auditor requires training in AI.". Then, the IT auditor will be guided with AI knowledge when performing the audit. Within the Audit controls component, information related to the current control will appear in the 'Guidance AI' column, if the IT auditor needs to audit AI-specific elements. By implementing this feature, the framework meets the artifact requirement of CA5.

The last question, question 5, is stated as follows: "Is the auditor well-informed about the field that the AI system is used in, including the relevant risks?", which can be answered with 'Yes' or 'No'. If this question is answered with 'Yes', the following text will appear in the right column: "The auditor does not require training in the field of [Answer question 1].". If question 5 is answered with 'No', the following action will appear in the action column: "The auditor requires training in the field of [Answer question 1].". This training is not included in the framework, so the IT auditor must follow this training within its own working time, before the audit starts.

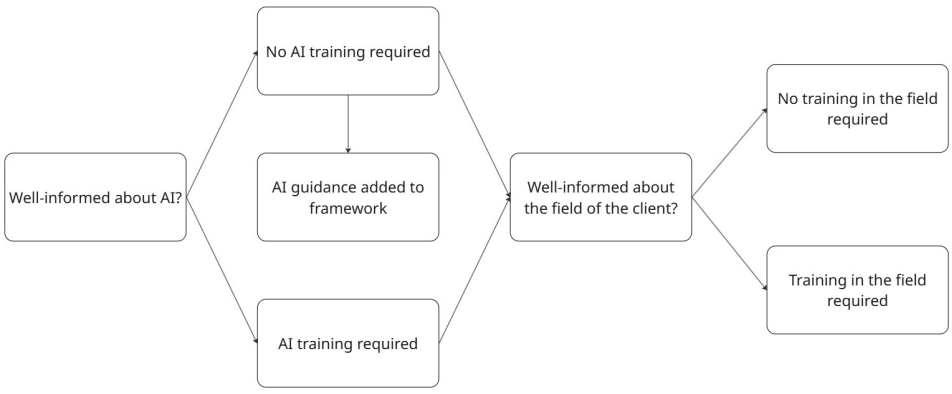



Figure 7.3: A visualization of the auditor requirements questions in the preparation process.

A visualization of the auditor requirements part is shown in figure 7.3. This question contributes partly to artifact requirement CA3, CA6, and CA7. When the IT auditor is well-informed about the field where in the AI system is used, it could assess the risks of a specific AI system in a specific sector more effectively. An IT auditor could measure bias and fairness in AI deployment within a specific sector more precisely if they have sufficient knowledge of that industry.

## 7.2 Audit controls

The second component, the Audit controls, is the main part of the AI audit framework and can be found in Appendix C. Here, the controls are described that the IT auditors need to test in order to determine compliance to the AIA. Therefore, IT auditors use this part of the framework while conducting the audit. The controls cover all the sub-articles of the AIA that are relevant to the scope of the audit, to determine compliance for a specific client. For this reason, our framework meets the artifact requirement of CA1. The controls are now focused on providers of AI-driven risk assessment models used in HLIC, as this is needed to test the framework on our mock demonstrator case.

The first and second columns of the Audit controls component contain the 'Control reference' and the 'Control name'. Every row is defined as a control, with a specific Control reference and Control name. Every control covers one or more sub-articles of the AIA, which are stated in the 'Article reference' column. Furthermore, the Annex number is added here, when it contains relevant information for the control. The next column is named 'Control procedure description' and contains detailed instructions for the IT auditor to test the control. For some controls, it is not specifically specified how the control should be tested. For instance, control 7.1 states that the auditor should assess whether the accuracy of the model is appropriate, but the instructions for testing this is lacking. This is due to the fact that the EU still needs to publish the required methods and accuracy levels. Therefore, the auditor is free to determine compliance based on professional judgment, which refers to the application of prior knowledge and experience to form audit conclusions [14]. Chiş et al. [14] argue that professional judgment is crucial to the performance of an audit, as it supports the interpretation of relevant ethical requirements and standards needed for the audit. In our framework, we emphasize that auditors need knowledge of current AI theory and practices, as well as the field in which the AI system is applied. This knowledge is essential, as it enhances the auditor's ability to interpret relevant ethical requirements and standards. Therefore, this knowledge strengthens the auditor's professional judgment. Considering the limited AI expertise among IT auditors and the lack of established best practices within AI auditing, the value of professional judgment can be questioned for our AI audit framework. However, due to the challenge of defining measurable criteria for some obligations, we believe that professional judgment remains the most effective way for auditing some controls under the AIA. Nevertheless, it is crucial that the IT auditor is sufficiently well-informed about current AI theory and practices, and the field in which the AI system is used.

Furthermore, the 'Needed audit evidence' column is added. This column contains all the evidence that the client needs to provide to the IT auditor, in order to test the specific control. For some controls, screenshots are part of the defined Needed audit evidence. IT auditors often need to test a control based on screenshots. For instance, control testing based on screenshots is commonly done within DSA and DMA audits. Therefore, IT auditors are familiar with this way of auditing. Besides this, Sandvig et al. [77] explain that using screenshots as evidence for algorithm auditing is a proper method. This therefore contributes to meeting the artifact requirement of CA2.

In the column right from 'Needed audit evidence', the 'Control guidance' is placed. In some these cells, extra instructions for the testing of the control is placed. This provides extra information about a control, for example, that a control only applies to testing datasets, if the client does not train the model itself. In addition, information in this column supports the auditor in measuring technical characteristics of the AI system, by giving examples of how this could be measured. Therefore, this column contributes to the artifact requirement of CA4 and CA5. Moreover, the artifact requirement of CA7 is met by implementing this column. This artifact requirement state that fairness, bias, and data drift within risk assessment models used by HLIC must be audited, when using the framework. One of the cells suggest to use Aequitas, which is a fairness assessment tool that can be used for audit [71]. The disparate impact ratio can be measured with this, for instance. This metric can be used to measure fairness within a model, as the ratio of positive outcomes of different subgroups are compared.

The following column is the 'Conclusion' column. Here, the IT auditor states its conclusion about the implementation of the control. There are 4 possible options that the IT auditor can choose from in the drop down menu. This scoring metric is chosen by comparing the audit scoring metrics of DSA audits that are already done. DSA audits are commonly done by the 'Big four' audit firms, which makes their scoring labels credible for usage within our framework. The DSA report of Tiktok performed by KPMG had the following three scoring labels: 'Positive', 'Positive with comments', and 'Negative' [44]. The DSA report by Meta, performed by the audit firm EY [22], used almost the same metric. This metric consists of the four labels: 'Positive'; 'Positive with comments'; 'Negative - partial non-compliance ('except for')'; 'Negative - full non-compliance ('adverse')'. We chose the last metric is to use in our framework, as the fourth label provides auditors the option to make a distinction between two types of negative control results. This metric contributes to CA4, as it makes the controls more measurable. The four scoring labels are described as follows:

- **Positive** (Green colored). The IT auditor finds that the provided evidence demonstrates that the client is compliant with the sub-article(s).
- **Positive with comments** (Yellow colored). The IT auditor finds that the client is mostly compliant with the sub-article(s). However, the IT auditor identified some weaknesses in the AI system

regarding this control.

- **Negative - except for** (Orange colored). The IT auditor finds that the client is mostly non-compliant with the sub-article(s). However, the IT auditor finds that a few parts of the evidence shows compliance with the sub-article(s).
- **Negative - adverse** (Red colored). The auditor finds that the client is non-compliant with the sub-article(s).

When the control conclusion of the IT auditor is stated as 'Positive with comments', 'Negative - except for', or 'Negative - adverse', it must fill in the cell next to the conclusion. This column is defined as 'Control gaps' and provides the IT auditor the space to explain what gaps exist in the control test. The gaps could consists of missing evidence, or evidence that is not sufficiently implemented or not implemented. The last column is the 'Guidance AI' column, which was already explained in the last subsection.

### 7.2.1 Summary Audit controls

To summarize, the Audit controls component is the main part of the AI audit framework. This component consists of practical controls for auditing providers of AI-driven risk assessment models used by HLIC. Each control contains a control reference and name, general control description, and article reference. Furthermore, it contains a precise control procedure description, the required audit evidence, and expended control guidance where needed. Then, the IT auditor is able to fill in the control conclusion and, where applicable, the control gap. The IT auditor is assisted with AI knowledge regarding AI-related controls, when the IT auditor is insufficiently informed about current AI theory and practices.

The controls are designed to cover all relevant high-risk obligations for providers of AI-driven risk assessment models used by HLIC. These obligations are related to risk management, data governance, technical documentation, record-keeping, transparency, human oversight, accuracy, robustness, cybersecurity, quality management, post-market monitoring, and serious incident reporting. For some controls, the IT auditor needs to rely on professional judgment. Although this may not be the most effective audit approach, we believe that it is necessary due to the lack of clarity within the AIA. This was a complicated trade-off, however, we made this choice consciously. We believe that the risks associated with professional judgment can be mitigated by training IT auditors in AI properly. Finally, the use of the four scoring labels creates a clear and standardized audit approach, while reducing ambiguity in audit outcomes.

# Chapter 8

# Evaluation

In this chapter, we present the evaluation phase of our research. First, we explain the selection of the interviewees. Then, we explain how the evaluation interviews were conducted. Lastly, we elaborate on the interview answers for each evaluating criterion.

## 8.1 Evaluation interviews

The evaluation process was conducted through interviews with five experts. These were almost the same participants who took part in the exploratory interviews, as described in Section 3.1.3. Only expert 1 did not participate in the evaluation interviews, because she had left KPMG. Therefore, we interviewed expert 6 instead. The profile of expert 6 is presented in table 8.1.

| | Role | Experience | Expertise |
|---|---|---|---|
| Expert 6 | Senior manager | ± 16 years | Responsible AI |

Table 8.1: The profile of expert 6.

At the beginning of each interview, we provided the expert with the synthetic data of NexaSure. After the expert had read these documents, we presented the AI audit framework to him or her. We explained the framework and gave the expert sufficient time to read this. Then, we asked the expert to imagine that he or she was auditing NexaSure under the AIA using the provided framework. Afterward, we asked the evaluation questions. To ensure the interviews align with the four criteria of Lincoln et al. [51], we designed the interviews as follows to meet each criterion:

- **Credibility**. To ensure credibility, all interviews included clarifying questions after each criteria scoring question. The experts were asked why they assigned a particular score to each criteria. If an expert had already provided an explanation when giving the score, the clarifying question was not asked.
- **Transferability**. To ensure transferability, all experts were asked to assess the usability of the framework.
- **Dependability**. To ensure dependability, all interviews were conducted separately. Furthermore, the interviews were semi-structured, which allowed us to ask follow-up questions. However, these follow-up questions were limited to clarifying the expert's previous response, to maintain consistency across interviews.
- **Confirmability**. To ensure confirmability, all interviews were recorded. This allowed us to capture all the explanations of the experts, which helped us in grounding our findings regarding the evaluation of our framework.

In addition to the usability topic, the interview questions included five other topics. These topics are relevance, completeness, understandability, logic, and added value. These topics were derived from the

method of Lincoln et al. [51]. We asked the experts to rate each topic on a scale ranging from 1 to 10. The questions for the evaluation interviews can be found in Appendix A.2.

| | Expert 2 | Expert 3 | Expert 4 | Expert 5 | Expert 6 | Average |
|---|---|---|---|---|---|---|
| Relevance | 9.0 | 9.0 | 7.0 | 7.0 | 10 | 8.4 |
| Usability | 7.5 | 8.0 | 6.0 | 6.5 | 7.0 | 7.0 |
| Completeness | 8.5 | 8.0 | 6.0 | 7.0 | 7.0 | 7.3 |
| Understandability | 8.0 | 9.0 | 5.0 | 6.0 | 8.0 | 7.2 |
| Logic | 7.0 | 8.0 | 7.0 | 6.0 | 9.0 | 7.4 |
| Added value | 7.5 | 9.0 | 6.0 | 6.5 | 7.0 | 7.2 |
| Average | 7.9 | 8.5 | 6.2 | 6.5 | 8.0 | |

Table 8.2: The evaluation scores of each category that were assigned by the experts.

In table 8.2, we present the evaluation scores that the experts assigned to each category. The last column shows the average score per category and the last row shows the average given score per expert. We elaborate on these scores and the explanations in the next sections.

### 8.1.1 Relevance

The experts were asked how relevant the AI audit framework is for third-party audit firms conducting audits under the AIA. The relevance category received the highest average score of all categories. The relevance scores are ranging from 7.0 to 10.0, which makes the range relatively large. Many experts emphasized the importance of the AIA and such a framework, as the demand for AI auditing increases and no such frameworks exist yet. Expert 3 explained during the interview:

> "I think it's quite relevant, given that we as auditors are being increasingly confronted with AI systems right now, on the bright side, and clients actively approach us, asking for help; Help auditing these systems themselves, or these systems are becoming a scope of the financial statement audit, for instance."

However, experts 4 and 5 were slightly less positive. They argued that the relevance is very situation dependent, as every audit is different. Therefore, the framework needs to be tailored to every specific situation to be relevant.

### 8.1.2 Usability

The experts were also asked to assess the usability of the framework. The experts were asked to imagine that they were auditing NexaSure under the AIA using the provided framework, so they had to assess the usability within this context. The usability category received the lowest average score of all categories. The range of the usability scores is 2.0, which is relatively low. Experts 3 explained that the basis of this framework was similar to that of other projects that it had worked on. Furthermore, expert 6 stated that the framework serves as a first step in translating the legal obligations into concrete control procedures. Expert 5 stated during the interview:

> "I think there is still improvement possible, but I think it's a good base."

Experts 4, 5 and 6 mentioned that certain details were missing, such as the frequency of the control, the risk that the control covers, and additional examples. Moreover, for expert 6, it was not clear when a certain control is tested positive, as it was not clear when the criteria is met. Lastly, expert 3 argued that the main question is whether it is feasible and recommended for the auditor to test the controls itself, or rely on controls implemented by the client.

### 8.1.3 Completeness

Furthermore, the experts were asked to evaluate the completeness of the framework. The completeness category received an average score of 7.3. The range of this category is average, as the ratings range from 6.0 to 8.5. Many experts mentioned that the framework was difficult to assess in terms of completeness, as they did not have detailed knowledge of the legal requirements of the AIA. Therefore, they were

uncertain whether all relevant controls had been included in the framework. Auditor 6 explained during the interview:

> "It looks structured and as if it could be complete, but the auditing answer is no, no basis to make any assertions about that."

However, experts 2 and 3 thought that the framework looked complete. As a result, these experts assigned relatively high ratings to the framework.

### 8.1.4 Understandability

In addition, the experts were asked to assess the understandability of the framework. They had to assess this from the perspective of an IT auditor, as they were asked to audit NexaSure using the framework. The understandability category received an average score of 7.2. This category has the highest range, as the scores range from 5.0 to 9.0. Experts 2, 3 and 6 could understand the framework relatively well. Expert 2 explained why he assigned an 8.0 to this category:

> "Because of the level of abstraction, it's nice and gives you a good overview."

On the other hand, expert 6 also emphasized the lack of details and concreteness within the controls. Experts 4 and 5 explained that the controls were not sufficiently clear, as the risks were missing and the criteria for a control to be compliant were unclear.

### 8.1.5 Logic

Then, the experts were questioned about their opinion regarding the logical structure of framework. The logic category received a average score of 7.4 and had a relatively high range of 3.0. Many experts thought that the framework was logically structured, partly due to the included preparation component. Expert 4 explained during the interview:

> "I do like the questions you raised in the first step, it does make sense, to determine the scope of the audit."

Experts 2 and 3 even mentioned that they would like to see more questions regarding the scope of the audit and the client. However, they would prefer fewer questions about auditor requirements, as they do not fit within an audit framework. Expert 5 believed that the logical structure was lacking, as the risks were not included. Lastly, experts 2 and 6 indicated that they missed a control mapping, which would create a clearer overview. The control mapping could be presented in the form of a menu, where the controls are mapped in themes.

### 8.1.6 Added value

Moreover, the experts were asked about the added value of the framework to the AI auditing ecosystem. For the added value category, the framework received an average score of 7.2. The range was relatively high, with scores ranging from 6.0 to 9.0 again. Experts 2, 3 and 6 believed that this framework could add value to the AI auditing ecosystem, mainly because there are currently few such frameworks available. Expert 2 explained during the interview:

> "So, in terms of doing a full audit, it wouldn't be complete, but it's definitely a good thing that helps or assists you during audits."

And expert 3 stated during the interview:

> "So, any contribution to do it in a more structured way is, I think, a welcome addition."

On the other hand, experts 4 and 5 raised some critical points regarding the framework's added value. They mentioned that the added value depends on the type of audit for which the framework is used. Moreover, the framework only adds value to the AI auditing ecosystem, if it is widely adopted by organizations and regulators. This is the case with DSA audits, for example.

### 8.1.7 Additional comments

Lastly, the experts were asked whether they had additional comments on the framework that were relevant to our research. Experts 2 and 3 had no additional comments.

Expert 4 emphasized that it is recommended to rely on the controls implemented by the client. Furthermore, it is important to set up benchmarks for the controls. By implementing this, the auditor knows when a certain system's characteristic is good enough during the control testing.

Expert 5 argued that the main thing that is missing in the framework, is the risk that each control covers. Moreover, expert 5 recognized the scoring labels from DSA reports, but suggested to look into the ISAE 3000 standard for this. Commonly used terminology for assurance engagements are described in the ISAE 3000, including scoring labels for controls.

Expert 6 noted that the control procedures need to be more concrete. References to industry standards, guidelines, and best practices would help IT auditors determine what evidence should be assessed at the client. Furthermore, such references would contribute to a more standardized audit approach and reduce reliance on professional judgment. However, expert 6 also noted that this remains challenging, as AI auditing is still a new field with limited established audit practices and benchmarks.

# Chapter 9

# Discussion

In this chapter, we present the discussion of our findings. We aim to interpret these findings and connect them to the literature. By explaining how and why the results came to be, we seek to provide a better understanding of auditing AI-driven risk assessment models under the AIA, which is the main objective of this research. Specifically, we dive into the IT auditor expertise, standardization versus sector-specific customization trade-off, relevant insurance audit topics, and the maturity of the AI auditing ecosystem. Lastly, we discuss the practical implications and limitations.

## 9.1 AI audit framework

First, we observed that the AIA contains many abstract requirements for high-risk AI systems, such as 'appropriate accuracy', 'effective human oversight', and 'acceptable risk levels'. Due to these vague specifications, it is complicated to translate the obligations into concrete audit procedures. In addition to this, the harmonized standards of the CEN/CENELEC JTC21, which should provide more guidance regarding AIA compliance, will probably be finalized in 2026. This finding is consistent with Ntaoulas et al. [63] and Sarkar et al. [78], who highlight the implementation gap between legal requirements and practical AI auditing. Moreover, the findings of the exploratory interviews showed that IT auditors face the challenge of operationalizing the AIA. The experts confirmed that translating the AIA obligations into measurable criteria is complicated, partly due to the abstract requirements for high-risk AI systems. This challenge directly influences the design of the AI audit framework, which seeks to translate these obligations into concrete and measurable controls.

After we designed the AI audit framework, we tested the framework through a mock demonstrator case and evaluation interviews. During the evaluation interviewees, the AI audit framework received relatively good scores, except for one 5.0 by expert 4. What we noticed was that the experts with RAI expertise assigned relatively high scores, compared to the experts with IT audit expertise. The average scores given by RAI experts were 7.9, 8.5 and 8.0, while the average scores given by IT audit experts were 6.2 and 6.5. This may be a result of their opinion and expertise regarding AI auditing. The RAI experts are constantly busy with AI-related work, which affects their opinion about the relevance in the first place. Secondly, this makes them relatively more suitable to rate the AI-related elements of the framework, compared to the IT audit experts. The IT audit experts have relatively more IT audit skills, knowledge and experience than the experts with RAI expertise. For this reason, the IT audit experts were more suitable to rate the audit elements specifically than the RAI experts.

Overall, the experts gave the framework a moderate score for understandability, but the differences between the ratings were relatively high, especially between the RAI experts and IT audit experts. This may be explained due their difference in expertise, as earlier explained, and by the fact that the experts interpreted the question differently. The IT audit experts assigned low ratings due to the limited level of detail in the framework. RAI experts assigned high ratings, as they appreciated the framework's level of abstraction and its ability to provide a clear overview of the audit process. This finding reflects the a challenge in AI auditing. Traditional IT auditing is supported by mature and detailed frameworks. However, for AI auditing, these mature and detailed frameworks have yet to be developed. Therefore, the IT audit experts have more audit frameworks against which they can compare our audit framework

with, making them more critical on its level of detail than the RAI experts.

The experts emphasized the relevance of the framework and considered it as a practical starting point for AI auditing under the AIA. Also, determining the audit scope in the Preparation process component was highly valued by the experts. Some experts even suggested to expand the scope questions, as it influences which risks, obligations, and controls are relevant for the audit. This implies that third-party audit firms need to devote significant attention to determining the scope of an AI audit. At the same time, the experts also indicated that the controls still lacked certain details, such as control frequencies, clearer criteria and risk mapping. It was complicated for us to implement these, as both the AIA and existing AI audit frameworks have similar shortcomings. This finding reflects the current AI auditing field, in which we know which elements are required for an AI audit procedure, but not how to formalize them precisely.

The challenge for the experts to assess the completeness of the framework was mainly due to their limited familiarity with the exact contents of the AIA. It was interesting to see that even the RAI experts were not fully familiar with the AIA, despite regularly working on AI-related projects. This suggests that third-party audit firms need to develop a clear understanding of the AIA contents, before designing and adopting an audit framework for the AIA.

### 9.1.1 IT auditor expertise

During the interviews, the experts were asked whether traditional IT auditors currently possess sufficient AI knowledge and skills to conduct high-quality AI audits and assessments. The general answer was that their current level of AI knowledge and skills is not yet sufficient. The traditional IT auditing skills alone are insufficient when auditing or assessing AI systems, as it is crucial to understand how an AI system works, how it processes data, and which risks it may introduce. This finding is consistent with Gadekallu et al. [27], who emphasize the lack of RAI knowledge within organizations and the need for RAI trainings. The finding is also consistent with Lam et al. [48], who identify the competence of an IT auditor as a precondition for effective AI auditing. This implies that AI auditing is not only an extension of IT auditing, but rather becoming a specialized discipline that requires technical and regulatory expertise. This aligns with the answers of expert 4, who explained that specialized AI profiles and expertise are needed for auditors in order to audit AI systems effectively. In addition, existing AI audit frameworks and the experts emphasize that AI auditing requires sector-specific methodologies. Since IT auditors must sometimes rely on professional judgment, they require knowledge regarding the sector that the audited AI system is used in.

For these reasons, we added questions regarding the IT auditor's expertise to the Preparation process component in the framework. These were questions about the AI and sector-specific knowledge of the IT auditor. The findings of the evaluation interviews show that these questions were not typically included in audit frameworks. Therefore, these should be removed when refining the framework. However, the experts still agreed that it is essential to check whether the IT auditor is well-informed about current AI theory and practices, and the field that the AI system is used in.

### 9.1.2 Standardization versus sector-specific customization

The literature review showed that various standards and frameworks already exist (ISO/IEC, IEEE, NIST, NOREA, KPMG, IIA, ForHumanity). However, none of these standards and frameworks specifically address third-party audits of AI-driven risk assessment models under the AIA, or high-risk AI systems in general. This finding supports the claim made by Mökander [56] that AI auditing methods lack standardization and mature practices. The existing standards and frameworks provide useful principles and guidance, but none of these are specifically meant for high-risk AI systems used in health and life insurance.

In our research, we tried to scope the audit framework specifically to providers of AI-driven risk assessment models used by HLIC. The evaluation interviews showed that experts viewed this as both a strength and a limitation. On the one hand, they valued the structured format of the framework, as it is relatively reusable for audits of providers of high-risk AI systems. On the other hand, they also emphasized that the framework would need to be tailored to audits with a different scope. This suggests that the AI auditing under the AIA needs a modular approach: a standard core of audit logic and controls, combined with sector-specific adjustments related to the relevant risks, obligations, and control

procedures.

### 9.1.3 Fairness, bias, and data drift

For AI-driven risk assessment models used in health and life insurance, the experts emphasized that fairness, biased data, and data drift require special attention. This finding is consistent with Nagamani [57] and Mitta [55], who highlight the ethical considerations and the need for regular audits of these models due to the risk of biased data. The predictions of these models can directly affect the underwriting decisions and insurance premiums. Even small biases in the underlying data or model behavior may lead to unfair disadvantages for certain groups or individuals. Therefore, it is essential to audit AI-driven risk assessment models used in health and life insurance with particular attention to fairness, bias, and data drift.

### 9.1.4 The maturity of the AI auditing ecosystem

For the exploratory interviews, the main objective was to gain insights into the mechanisms currently used by third-party auditors to assess AI systems, as well as the requirements and specifications that should be applied when auditing AI systems. The interviews were conducted with experts of KPMG, which is why we can only draw conclusions about the practices and perspectives of KPMG experts. The first thing we noticed was that currently AI audits are not often conducted by IT auditors. This is partly due to the fact that an audit by an independent auditor is not required by the AIA. However, IT auditors have received requests from internal auditors to support them with AI audits. Apart from these engagements, only assessments have been conducted by IT auditors. Some of these were readiness assessments, where IT auditors determined whether the client was ready for the AIA. In addition, risk assessments have been conducted by IT auditors, in which they assessed the risks of AI systems and provided advice on how to mitigate these. Overall, the experts stated that the demand for AI audits and assessments is currently increasing. However, they stated that there is no shared framework that can be used for these audits and assessments. Most of the time, IT auditors try to combine different parts of existing frameworks and use that for the audit or assessment. This finding aligns with the findings of Mökander [56], as they highlight the lack of shared frameworks for AI audits.

What is interesting is that, despite the diversity of standards and frameworks, almost all of them emphasize similar principles, such as transparency, data governance, fairness, security, accountability, and accuracy. These principles align with the themes of the high-risk obligations of the AIA. This suggests that a consensus is emerging between organizations, regulators, and bodies that develop standards, regarding what responsible AI means and what requirements AI systems should meet. Nevertheless, the audit methodologies for responsible AI and AIA compliance remain underdeveloped.

Furthermore, the experts explained that challenges sometimes arise when obtaining evidence from the client for the audit or assessment. The experts indicated that clients generally provide more evidence during assessments than during audits, as assessments are often initiated by the clients themselves, whereas audits are often mandatory. As a result, clients are often more willing to share information during assessments because they aim to gain insights into their own level of compliance. This makes it easier for auditors to obtain the necessary evidence for the engagement.

In addition, experts noticed that the type of client affects both the availability of evidence and the collaboration process during an audit. According to the experts, online platforms and AI vendors are generally less used to audits than financial companies and consist of relatively many engineers. These engineers prefer to spend time on their own work rather than supporting IT auditors with their audit activities. Due to the limited audit knowledge of online platforms and AI-vendors, documentation that is required for the audit is often lacking. This is a challenge for both IT audit and AI audit, as it requires much time and effort to obtain the needed evidence for an audit. However, the audit maturity of these organizations is expected to increase over time, according to expert 4. Nevertheless, even when large online platforms and AI vendors are used to audits, they prefer to disclose as little evidence as possible. This evidence may contain sensitive and confidential information regarding their AI training techniques, as expert 3 explained.

These findings align with Raji et al. [74], who argue that external auditors often lack access to internal datasets, training procedures and development documentation. Together with the abstract obligations in the AIA, this may pose a challenge for the future of AI auditing under the AIA. For both auditors

and organizations, it would be complicated to determine what evidence is really necessary for the audit. Therefore, it is crucial that the AI audit framework clearly specifies which evidence should be obtained from clients.

The experts believed that the AI audit framework could add value to the AI auditing ecosystem, but it requires some adjustments and a widespread adoption by organizations and regulators. Nevertheless, the experts considered the framework a proper basis, as it translates the obligations into controls in a structured way. Given the fact that this was not done before, the experts highly valued the framework. This reflects the challenges that are described by Mökander [56], Ntaoulas et al. [63] and Sarkar et al. [78], who highlight the implementation gap between legal requirements and practical auditing methods.

Additional comments that were made by the experts included the observation that IT auditors could rely on controls implemented by the client. Although this point was valid, we question whether clients are sufficiently capable of implementing the controls themselves. Different interpretations of the AIA obligations may result in ineffective controls implemented by organizations. Moreover, an additional comment was that the concreteness of control procedures can be improved by adding references to industry standards, guidelines, and best practices. This would help to determine what evidence is needed, create a standardized audit approach and reduce reliance on professional judgment. We believe that these two comments may be addressed once the harmonized standards of the CEN/CENELEC JTC21 are published. The guidelines in these standards will provide clarity to organizations about when compliance is achieved, which supports them in implementing correct and effective controls.

## 9.2 Practical implications

This research has several practical implications for third-party audit firms, regulators, and providers of high-risk AI systems. First, for third-party audit firms, the findings show that AI auditing under the AIA requires a more standardized method, but at the same time also a sector-specific approach. Currently, AI assessments are often performed by combining different standards and frameworks, because no shared methodology exists yet. The AI audit framework developed in this research provides auditors with a practical starting point for translating the high-risk obligations of the AIA into concrete audit controls. Furthermore, the findings emphasize that auditors require both knowledge of AI and of the sector in which the AI system operates. Although this part should not be included of the audit framework, third-party audit firms should invest more in AI training programs and multidisciplinary audit teams to resolve this challenge.

Furthermore, the findings are relevant for regulators and standardization bodies. Both the exploratory and evaluation interviews showed that further guidance of the EU and standardization bodies are necessary. It is beneficial to the AI auditing ecosystem if the EU introduces an obligation on audits under the AIA, for providers and deployers of high-risk AI systems. Furthermore, the findings show that it is important that the harmonized standards of the CEN/CENELEC JTC21 will be published soon. These standards and audit obligation may help create consistency between organizations and auditors and reduce reliance on professional judgment.

Lastly, the results are relevant for providers of high-risk AI systems. The framework provides the first insights into the types of controls, documentation, and evidence that may be required during future audits under the AIA. Providers of high-risk AI systems can use this framework to assess their readiness for both AI auditing and the AIA. Therefore, the framework can be seen as a first guideline to prepare them for audits under the AIA. For providers, this would reduce the effort that is required to demonstrate compliance with the AIA. Especially for providers of AI-driven risk assessment models used by HLIC, the results of this research are relevant, as they show the relevant elements for their audit. Fairness, biased data and data drift are critical issues that require attention when auditing AI-driven risk assessment models used by HLIC under the AIA.

## 9.3 Limitations

In this section, we present the limitations of this thesis. A few limitations exist in our research, such as the small sample size, single case study, evolving regulation, and rater effect.

First, the small sample size limits the external validity of the interview findings. This means that it is unclear to what extent these findings can be generalized beyond the scope of this study. We conducted

both rounds of interview with five employees from the same department within one organization, namely the ITAA department of KPMG. Consequently, the findings of these interviews only reflect the perspectives of ITAA employees of KPMG and may not be representative of the general third-party audit profession.

Secondly, we conducted this research with a mock demonstrator case that included only one fictive case. This case was about a provider of an AI-driven risk assessment model used by HLIC. By focusing on a single case, we cannot draw conclusions about the specific requirements of other types of audits involving different AI systems, roles, and sectors. Nevertheless, we may state that some parts of the AI audit framework are reusable for other audits of providers of high-risk AI systems under the AIA. The structure of the audit framework, consisting of the Preparation process and Audit controls, can be reused for auditing providers of high-risk AI systems. The controls cover the relevant articles for these audits, however, the conformity assessments for AI systems used in biometrics or as safety components of products must be conducted differently. Therefore, control 12.1 of our AI audit framework cannot be used for auditing these types of AI systems. Furthermore, the Control procedure descriptions and any additional Control guidance cannot be reused for other types of AI audits. This is because some of these elements are specifically created for risk assessment models used by HLIC. Examples include the types of data that is allowed for model training, the fairness metric tool, and the data drift metrics. While some of these elements may be reusable for other types of high-risk AI systems, they are not relevant for all of them. Moreover, the AI audit framework is not reusable for auditing deployers, importers, or distributors of high-risk AI systems.

Moreover, the evolving regulatory landscape around the AIA is a limitation to our research. The Digital Omnibus on AI Proposal and the future guidelines of the CEN/CENELEC JTC21 harmonized standards may introduce significant changes to the current AI auditing ecosystem. Obligations, criteria, and deadlines for compliance may change in the future, which influences the requirements and relevance of the AI audit framework. Therefore, it is possible that certain results and findings of this research may become less applicable or relevant in the future.

Lastly, there is a possibility that the scores from the evaluation interviews are less reliable due to the rater effect. The rater effect consist of systematic biases or tendencies in how evaluators score performance, causing test results to reflect influences that are unrelated to the skill or ability that is measured [12]. One of these tendencies is the leniency bias, which is the tendency to give ratings that are more positive than justified. We believe that expert 6 may have been affected by leniency bias based on the answers that were given. Expert 6 assigned a 7.0 to completeness, although he stated that from an IT auditor perspective, the framework was not complete. This inconsistency suggests that expert 6 may have had a tendency to give higher ratings than justified.

# Chapter 10

# Conclusions

In this chapter, we present the conclusions of the research. First, we provide the answers to the sub-questions, followed by the answer to the research question. Furthermore, we elaborate on the contributions and future work of this thesis.

## 10.1 Answers to the research questions

This thesis investigated how an AI audit framework used by a third-party audit firm can be improved to effectively audit AI-driven risk assessment models used by HLIC under the AIA. To answer this question, we conducted a literature review and exploratory interviews, and we designed an upgraded AI audit framework. Then, we tested this framework through a mock demonstrator case and evaluated it through expert interviews.

**SQ.1** *What mechanisms do third-party audit firms currently use to assess AI systems?*

The findings show that third-party audit firms currently do not conduct audits that are only focused on AI system, but it is rather a part of the audit. Engagements that are only focused on AI systems are conducted as assessments, which can be readiness assessments or risk assessments. In AI-related engagements, IT auditors combine elements from existing standards and frameworks, such as from ISO/IEC, IEEE, NIST, NOREA, KPMG, IIA, and ForHumanity. Furthermore, during AI audits and assessments, IT auditors sometimes struggle with obtaining the right evidence. Also, they depend relatively much on professional judgment, as concrete criteria does not exist yet. However, most IT auditors do not have sufficient AI expertise to audit AI systems under the AIA, which makes this a challenge for the AI auditing ecosystem. In addition, the type of engagement and the type of client affects both the availability of evidence and the collaboration process during the engagement. Therefore, we conclude that current mechanisms used by third-party audit firms that assess AI systems are fragmented and still developing.

**SQ.2** *Following the adoption of the AI Act, what specifications and requirements must third-party auditors apply when auditing AI-driven risk assessment models used in health and life insurance?*

AI-driven risk assessment models used by HLIC fall under the high-risk category of the AIA when they are used for risk assessment and pricing in relation to natural persons. Therefore, third-party auditors need to assess whether providers of these systems comply with the relevant high-risk obligations of the AIA. These obligations are related to risk management, data governance, technical documentation, record-keeping, transparency, human oversight, accuracy, robustness, cybersecurity, quality management, post-market monitoring, and serious incident reporting. Moreover, the exploratory interviews showed that IT auditors should pay specific attention to fairness, bias, and data drift, as these aspects are especially relevant for AI-driven risk assessment models used by HLIC.

However, many obligations in the AIA are abstract and contain terms such as 'appropriate', 'effective', and 'acceptable'. This makes it difficult for IT auditors to determine when requirements are met. Therefore, a requirement for the AI audit framework is that these obligations are translated into measurable controls and practical audit procedures. Interestingly, the findings of the exploratory interviews show that the AI audit framework needs to be structured and standardized, but at the same time also sector-

specific. Lastly, we conclude that AI auditing is not only an extension of IT auditing, but rather becoming a specialized discipline that requires technical and regulatory expertise.

**SQ.3** *To what extent can an upgraded AI audit framework fulfill these specifications and requirements when tested through a mock demonstrator case?*

The upgraded AI audit framework was able to fulfill a significant part of the specifications and requirements when tested through the NexaSure mock demonstrator case. The framework consists of two components, which are the preparation component and the audit controls component. The preparation component contains a checklist to determine the scope and check whether the auditors have sufficient AI and sector-specific expertise. The audit controls component then translates relevant high-risk obligations into control procedures, including the required evidence, AI guidance, and conclusion labels. However, some specifications and requirements were hard to implement, such as measurable audit criteria. Therefore, we conclude that the upgraded AI audit framework can fulfill the specifications and requirements to a considerable extent, but further development is needed to properly implement them all.

**SQ.4** *What is the assessment of third-party auditors regarding the upgraded AI audit framework, as illustrated through a mock demonstrator case?*

The evaluation interviews showed that the experts considered the upgraded AI audit framework highly relevant for third-party audit firms that audit AI systems under the AIA. The framework scored well above sufficient on completeness, logic, understandability, usability, and added value. The expert especially valued the fact that the framework is a practical starting point for translating the AIA obligations into structured control procedures. Moreover, the experts highly valued the questions about the scope in the preparation component, but valued the questions about the auditor requirements less. The experts also identified other areas for improvements. The controls often lack certain details, such as clearer control criteria, risk mapping, control frequencies, and references to industry standards or best practices. The experts also emphasized that the framework needs to be tailored to the specific AI system, sector, and client role. Overall, we conclude that the upgraded AI audit framework was assessed as a relevant and promising first step, but not yet as a fully developed audit methodology.

**RQ.** *How can an AI audit framework used by a third-party audit firm be improved to effectively audit AI-driven risk assessment models used in health and life insurance under the AI Act?*

Based on the findings, we conclude that the AI audit framework can be improved by transforming the high-risk obligations of the AIA into a structured set of audit controls that include control procedures, evidence requirements, AI guidance, and conclusion labels. The framework should also include a preparation component in which the engagement lead determines the audit scope, including the client role and the risk classification of the AI system. For AI-driven risk assessment models used in health and life insurance, the framework should specifically include controls related to fairness, bias, and data drift, along with the high-risk obligations of the AIA. This thesis shows that such a framework can support third-party auditors when auditing providers of high-risk AI systems under the AIA, as the main parts of the framework are reusable for these audits. However, some elements of the framework should be tailored to the specific scope of the audit. Finally, the AI audit framework should be seen as a first version rather than a final audit methodology. Its effectiveness depends on further refinement, validation in additional cases, and alignment with future harmonized standards and regulatory guidance.

## 10.2 Contributions

This thesis makes two contributions. First, this thesis contributes to academic literature, as there is very few research done about auditing AI-driven risk assessment models used in health and life insurance under the AIA. Existing literature discusses AI auditing and AIA compliance, but limited research combine this with a specific case study to design an auditing framework for a particular AI system. This research therefore contributes to this gap by identifying relevant requirements and translating them into an audit framework.

Second, this thesis contributes to practice by providing third-party audit firms with a practical starting point for AI audits under the AIA. The upgraded AI audit framework helps IT auditors to create a structured audit methodology. This is especially relevant because the findings show that current AI audit practices and methodologies are fragmented and still developing.

## 10.3 Future work

In this section, we present several opportunities for future research. First, the AI audit framework should be validated by a larger and more diverse group of experts. Future studies can include IT auditors of other third-party audit firms. By evaluating the framework by these experts, a better understanding of the quality of the framework can be obtained.

Moreover, future research should include use cases of different AI systems, for which an AI audit framework under the AIA can be designed. This thesis included a single mock demonstrator case involving a provider of an AI-driven risk assessment model used by HLIC. However, if an AI audit framework would be designed for a different AI system, comparisons can be made to determine which parts are generally applicable and which parts require sector-specific customization.

Furthermore, future work should include the refinement of the upgraded AI audit framework. The experts explained that the framework requires more details in order to be complete and usable. This can be done by the IT auditors of KPMG, for example.

Lastly, the AI audit framework should be updated again after the publishing of the CEN/CENELEC JTC21 harmonized standards and the results of the Digital Omnibus on AI Proposal are clear. The requirements and timelines for AIA compliance may change due to these developments, and the requirements of the AI audit framework may therefore change as well.

# Bibliography


[1] GDPR Advisor. The 7 principles of gdpr, 2026. URL https://www.gdpr-advisor.com/the-7-principles-of-gdpr/.

[2] AP. The gdpr in brief, 2025. URL https://www.autoriteitpersoonsgegevens.nl/en/themes/basic-gdpr/gdpr-basics/the-gdpr-in-brief.

[3] AP. Zorgverzekeraars, 2026. URL https://www.autoriteitpersoonsgegevens.nl/themas/gezondheid/zorgverzekeraars.

[4] Zahed Ashkara. Digital omnibus & ai act: what changes for compliance, 2026. URL https://www.aiactblog.nl/en/posts/digital-omnibus-ai-act-simplification-or-erosion.

[5] IEEE Standards Association. Ieee standards association statement of intention our role in addressing ethical considerations of autonomous and intelligent systems (a/is), 2016. URL https://standards.ieee.org/wp-content/uploads/import/documents/other/ethical-considerations-ai-as-29mar2018.pdf.

[6] G. Barham, A. Banzon, A. Mercer, K. Seeuws, and G. Nordhoff. *THE IIA'S Artificial Intelligence Auditing Framework*, 2024. URL https://storage.ghost.io/c/44/95/449506ca-034e-480f-9725-fcde08ef1cc1/content/files/2025/05/Artificial-Intelligence-Auditing-Framework.pdf?ref=aigl.blog.

[7] Mariano Martín Zamorano Barrios. Ai audits: How do you implement the eu ai act?, 2024. URL https://trilateralresearch.com/artificial-intelligence/ai-audits-how-do-you-implement-the-eu-ai-act.

[8] A. Brackey. Analysis of racial bias in northpointe's compas algorithm. *ProQuest Dissertations and Theses*, page 46, 2019.

[9] D. Brand, M. Hoffmann, and J. Van der Merwe. Development of effective methods and tools for the auditing of ai algorithms by the supreme audit institutions. *eJournal of eDemocracy and Open Government*, 7:106–140, 2025.

[10] M. Breslin and R. Buchanan. On the case study method of research and teaching in design. *Design Issues*, 24(1):36–40, 2008.

[11] CEN-CENELEC. Cen and cenelec, 2026. URL https://www.cencenelec.eu/european-standardization/cen-and-cenelec/.

[12] K. H. C. Cheng, C. H. Hui, and W. F. Cascio. Leniency bias in performance ratings: The big-five correlates. *Frontiers in Psychology*, 8(521), 2017.

[13] D. Chicco, M. J. Warrens, and G. Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 2021.

[14] A. O. Chiş and A. M. Achim. Professional judgement. the key to a successful audit. *SEA: Practical Application of Science*, 2(3):217, 2014.

[15] The EU Artificial Intelligence (AI) Act: A Commentary. *N. Forgó and C. N. Pehlivan and P. Valcke.* Wolters Kluwer, 2024.

[16] THE EU REGULATION ON ARTIFICIAL INTELLIGENCE: A COMMENTARY. *A. J. Huergo Lora and G. M. Díaz González*. Wolters Kluwer, 2025.

[17] European Commission. *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI)*, 2025. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52025PC0836.

[18] European Commission. Ai act, 2026. URL https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

[19] J. Cowls, A. Tsamados, M. Taddeo, and L. Floridi. The ai gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendation. *AI & SOCIETY*, 38: 283—-307, 2023.

[20] M. de Boer and H. van Geijn. Norea guiding principles trustworthy ai investigations. *NOREA*, 2021.

[21] Euronews. Eu standards bodies flag delays to work on ai act, 2025. URL https://www.euronews.com/next/2025/04/16/eu-standards-bodies-flag-delays-to-work-on-ai-act.

[22] EY. *Independent Audit on Facebook*, 2024. URL https://panoptykon.org/sites/default/files/2024-11/ocena-ryzyka-dsa-2024_facebook.pdf.

[23] eyreACT. The eu digital omnibus explained: What it means for eu ai act enforcement dates in 2026, 2026. URL https://eyreact.com/the-eu-digital-omnibus-explained-what-it-means-for-eu-ai-act-enforcement-dates-in-2026/.

[24] Directorate-General for Communication. Ai act enters into force, 2024. URL https://commission.europa.eu/news-and-media/news/ai-act-enters-force-2024-08-01_en.

[25] ForHumanity. Independent audit of ai systems, 2016. URL https://forhumanity.center/independent-audit-of-ai-systems/.

[26] ForHumanity. *EU AI Act - Provider only*, 2024. URL https://forhumanity.center/site/wp-content/uploads/2025/03/Excerpt-EU-Artificial-Intelligence-Act-Provider-only-v1.5.pdf.

[27] T. R. Gadekallu, K. Dev, S. A. Khowaja, W. Wang, H. Feng, K. Fang, S. Pandya, and W. Wang. Framework, standards, applications and best practices of responsible ai : A comprehensive survey. *Elsevier*, 2025.

[28] V. Garousi, M. Felderer, M. V. Mäntylä, and A. Rainer. Benefitting from the grey literature in software engineering research. *Contemporary Empirical Methods in Software Engineering*, pages pp 385—-413, 2020.

[29] S. Goellner, M. Tropmann-Frick, and B. Brumen. Responsible artificial intelligence: A structured literature review. 2024. URL https://arxiv.org/abs/2403.06910.

[30] E. G. González and P. de Hert. Understanding the legal provisions that allow processing and profiling of personal data—an analysis of gdpr provisions and principles. *ERA Forum*, 19:597—-621, 2019.

[31] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 2018.

[32] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain. Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16:45—-74, 2024.

[33] Maximilian Henning. High-risk ai guidelines will be late again, commission confirms, 2026. URL https://www.euractiv.com/news/high-risk-ai-guidelines-will-be-late-again-commission-confirms/.

[34] IEC. Welcome to the iec webstore, 2026. urlhttps://webstore.iec.ch/en/.

[35] IEEE. Ieee, 2026. URL https://www.ieee.org/.

[36] ISO. Iso: Global standards for trusted goods and services, 2026. URL https://www.iso.org/home.html.

[37] ISO/IEC. Iso 19011:2018 guidelines for auditing management systems. *International Standard*, (3), 2018.

[38] ISO/IEC. Iso/iec 38507:2022 information technology — governance of it — governance implications of the use of artificial intelligence by organizations. *International Standard*, (1), 2022.

[39] ISO/IEC. Iso/iec 23894:2023 information technology — artificial intelligence — guidance on risk management. *International Standard*, (1), 2023.

[40] ISO/IEC. Iso/iec 42005:2025 information technology — artificial intelligence (ai) — ai system impact assessment. *International Standard*, (1), 2025.

[41] A. Koshiyama, E. Kazim, P. Treleaven, P. Rai, L. Szpruch, G. Pavey, G. Ahamat, F. Leutner, R. Goebel, A. Knight, J. Adams, C. Hitrova, J. Barnett, P. Nachev, D. Barber, T. Chamorro-Premuzic, K. Klemmer, M. Gregorovic, S. Khan, E. Lomas, A. Hilliard, and S. Chatterjee. Towards algorithm auditing: managing legal, ethical and technological risks of ai, ml and associated algorithms. *Royal Society Open Science*, 11(5), 2024.

[42] KPMG. *The KPMG Trusted AI approach*, 2023. URL https://assets.kpmg.com/content/dam/kpmgsites/xx/pdf/2023/12/kpmg-trusted-ai-approach.pdf?v=latest.

[43] KPMG. Auditing artificial intelligence, 2025. URL https://kpmg.com/kpmg-us/content/dam/kpmg/pdf/2025/auditing-artificial-intelligence.pdf.

[44] KPMG. *DSA Assurance Report*, 2025. URL https://sf16-va.tiktokcdn.com/obj/eden-va2/zayvwlY_fjulyhwzuhy[/ljhwZthlaukjlkulzlp/DSA/DSAAssuranceReport-TikTokTechnologyLimited-2025.pdf.

[45] KPMG. Audit & assurance, 2026. URL https://kpmg.com/nl/en/services/audit-and-assurance.html.

[46] KPMG. Kpmg us, 2026. URL https://kpmg.com/us/en.html.

[47] I. Kusmaryono, D. Wijayanti, and H. R. Maharani. Number of response options, reliability, validity, and potential bias in the use of the likert scale education and social science research: A literature review. *International Journal of Educational Methodology*, 8(4), 2022.

[48] K. Lam, B. Lange, B. Blili-Hamelin, J. Davidovic, S. Brown, and A. Hasan. A framework for assurance audits of algorithmic systems. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1078—-1092, 2024.

[49] Raimund Laqua. Audits vs. assessments: Understanding the key differences, 2026. URL https://www.leancompliance.ca/post/audits-vs-assessments-understanding-the-key-differences.

[50] Y. Li and S. Goel. Making it possible for the auditing of ai: A systematic review of ai audits and ai auditability. *Information Systems Frontiers*, 27:1121—-1151, 2024.

[51] Y. S. Lincoln and E. G. Guba. *Naturalistic Inquiry*. Sage Publications, 1985.

[52] J. Liu, J. Lee, and R. Zhou. Review of big-data and ai application in typhoon-related disaster risk early warning in typhoon committee region. *Tropical Cyclone Research and Review*, 12(4):341–353, 2023.

[53] B. Lund, Z. Orhan, N. R. Mannuru, R. V. K. Bevara, B. P., M. K. Vinaih, and P. Bhaskara. Standards, frameworks, and legislation for artificial intelligence (ai) transparency. *AI and Ethics*, 5: 3639—3655, 2025.

[54] S. B. Mannapur. Understanding data drift and concept drift in machine learning system. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 11(1):318—-330, 2025.

[55] N. R. Mitta. Ai-based predictive analytics for life insurance underwriting: Leveraging machine learning models for mortality risk assessment, policyholder profiling, and premium calculation. *American Journal of Data Science and Artificial Intelligence Innovations*, 2:327—362, 2022.

[56] J. Mökander. Auditing of ai: Legal, ethical and technical approaches. *Digital Society*, 2(49), 2023.

[57] N. Nagamani. Ai-driven risk assessment models for health and life insurance underwriting. *International Journal of Computer Science and Engineering*, 5(1), 2025.

[58] NIST. Artificial intelligence risk management framework (ai rmf 1.0). *NIST*, 2023. URL `https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf`.

[59] NIST. National institute of standards and technology, 2026. URL `https://www.nist.gov/`.

[60] NOREA. Over norea, 2026. URL `https://www.norea.nl/over-norea`.

[61] C. Novelli, F. Casolari, A. Rotolo, M. Taddeo, and L. Floridi. Ai risk assessment: A scenario-based, proportional methodology for the ai act. *Digital Society*, 3(13), 2024.

[62] C. Novelli, P. Hacker, J. Morley, J. Trondal, and L. Floridi. A robust governance for the ai act: Ai office, ai board, scientific panel, and national authorities. *European Journal of Risk Regulation*, 16 (2):566—590, 2025.

[63] V. Ntaoulas, K. Mpouros, V. Galiatsatou, A. Tsakiri, T. Giavoudi, and V. Manousopoulos I. Krasonikolakis. Towards accountable ai: Implementing risk-based audits under the eu ai act. *18th Annual Conference of the EuroMed Academy of Business*, 2025.

[64] Future of Life Institute. The ai act explorer, 2024. URL `https://artificialintelligenceact.eu/ai-act-explorer/`.

[65] Future of Life Institute. Standard setting, 2025. URL `https://artificialintelligenceact.eu/standard-setting-overview/`.

[66] Official Journal of the European Union. *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, 2016. URL `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679`.

[67] Official Journal of the European Union. *Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act)*, 2022. URL `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R1925`.

[68] Official Journal of the European Union. *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)*, 2022. URL `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2065`.

[69] Official Journal of the European Union. *REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act*, 2024. URL `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689`.

[70] O. S. Onwuajuese, H. Rafiq, S. McHale, P. O. Ugochukwu, and S. Hunter-Barnett. Ai-driven risk assessments: Advancing cybersecurity and sustainability. *Cybersecurity and Human Capabilities Through Symbiotic Artificial Intelligence*, pages 327—337, 2025.

[71] C. S. Pareek. Unmasking bias: A framework for testing and mitigating ai bias in insurance underwriting models. *Journal of Artificial Intelligence, Machine Learning and Data Science*, 1(1):1–7, 2023.

[72] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee. A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77, 2007.

[73] P. J. Phillips. Four principles of explainable artificial intelligence. *NIST*, 2021.

[74] I. Raji, A. Smart, R. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 33—44, 2020.

[75] N. Razavian, F. Knoll, and K. J. Geras. Artificial intelligence explained for nonexperts. *Seminars in musculoskeletal radiology*, 24(1):3–11, 2020.

[76] M. A. Rothstein. *Genetics and life insurance : medical underwriting and social policy*. MIT Press, 2004.

[77] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, 2014.

[78] S. Sarkar, N. Sunheriya, J. Giri, K. Al-Qawasmi, and R. Chadge. A comprehensive quantitative model for ethical ai risk assessment: Eu act on artificial intelligence. *Artificial Intelligence in the Digital Era*, pages 145—165, 2025.

[79] M. van Bekkum, F. Z. Borgesius, and T. Heskes. Ai, insurance, discrimination and unfair differentiation: an overview and research agenda. *Law, Innovation and Technology*, 17(1), 2025.

[80] verbond van verzekeraars. *GEDRAGSCODE VERWERKING PERSOONSGEGEVENS VERZEKERAARS*, 2024. URL `https://www.verzekeraars.nl/media/3lpdl0lx/opgemaakte-definitieve-versie-gvpv-2024.pdf`.

[81] W. Wang and Y. Lu. Analysis of the mean absolute error (mae) and the root mean square error (rmse) in assessing rounding model. *IOP Conference Series: Materials Science and Engineering*, 2018.

[82] R. J. Wieringa. *Design Science Methodology for Information Systems and Software Engineering*. Springer Berlin, Heidelberg, 2014.

[83] Ben Wolford. What is gdpr, the eu's new data protection law?, 2026. URL `https://gdpr.eu/what-is-gdpr/`.

[84] B. Yurdakul. Statistical properties of population stability index. *ProQuest LLC*, 2018.

# Appendix A

# Interviews

In this section, we present the questions of the exploratory and evaluation interviews. After that, the thematic coding is shown.

## A.1 Exploratory interview questions

The interview questions were developed based on the problem–policy–impact framework. This framework enables a systematic understanding of (1) bottlenecks in the current practice of AI auditing, (2) existing and missing policy instruments such as frameworks and legislation, and (3) the importance and impact of AI auditing.

**Introduction**

1. What is your role within KPMG, and what does your work involve?

**A. Problem framework – current practice and bottlenecks**

2. To what extent are AI audits or assessments currently being conducted within ITAA, and how does such a process typically unfold?

3. What bottlenecks or challenges is ITAA currently experiencing when conducting AI audits or AI assessments?

4. How would you assess the current auditability of clients' AI systems? Do they provide sufficient documentation for an AI audit or assessment, and is there adequate transparency regarding the processes they perform?

5. Do you think IT auditors currently have sufficient AI knowledge and skills to conduct high-quality AI audits and assessments?

**B. Policy framework – rules, frameworks, and guidelines**

6. What is your opinion on the frameworks and guidelines currently used for auditing and assessing AI systems?

7. Do you believe these frameworks and guidelines are sufficient for the future of AI auditing under the AI Act?

8. What, if anything, is missing from these frameworks and guidelines?

9. What additional considerations should be taken into account when auditing risk assessment models used by health and life insurance companies under the AI Act?

**C. Impact framework – consequences, impact, and future outlook**

10. To what extent do you expect the demand for AI audits to change in the future, given that the AI Act does not impose a direct audit obligation?

11. What risks do you believe arise when AI auditing is insufficiently developed or inadequately applied?

12. What additional changes are needed to stimulate a well-functioning AI auditing ecosystem? For example, are organizational or legal adjustments required?

## A.2 Evaluation interview questions

1a. On a scale of 1 to 10, how relevant is this AI audit framework for third-party audit firms conducting audits under the AI Act?

1b. Why would you give this rate?

2a. On a scale of 1 to 10, how would you rate the usability of this AI audit framework?

2b. Why would you give this rate?

3a. On a scale of 1 to 10, how would you rate the completeness of this AI audit framework?

3b. Why would you give this rate?

4a. On a scale of 1 to 10, how would you rate the understandability of this AI audit framework?

4b. Why would you give this rate?

5a. On a scale of 1 to 10, how would you rate the logical structure of this AI audit framework?

5b. Why would you give this rate?

6a. On a scale of 1 to 10, how would you rate the added value of this AI audit framework to the AI auditing ecosystem?

6b. Why would you give this rate?

7. Do you have any other comments on this framework that are relevant to my research?

## A.3 Thematic coding

In this section, we present the data matrix of the thematic coding.

| | | | Codes | | | | | |
|---|---|---|---|---|---|---|---|---|
| Interviewees | Job title | Experience | AI audits done | AI audit process | Challenges | Auditability | AI skills | Frameworks existence |
| Expert 1 | Consultant RAI | 1 year | AI audits are not taking place currently, but I helped an internal auditor with an AI audit once. | With DSA/DMA audits, AI is partly audited on these laws. | AI auditing within ITAA is quite new. There is no standardized method for AI auditing, which results in auditors doing it on their own way. | For an audit, clients only provide required evidence but nothing more, because it is confidential. For readiness assessments, clients request this themselves and are willing to provide a lot of evidence. | IT auditors don't have enough AI skills and knowledge to audit AI system, but they can learn it. | ISO standards exist regarding risk management and AI systems, but is not used for audits. I combined different aspects from different frameworks when I did that AI audit, because there is no obligation how this should look like. |
| Expert 2 | Manager RAI | 7 year | Not much AI audits, but it increases slowly. | When ITAA audit involves AI, we help them. Or we help internal auditors with AI audits. You define the scope, test against a framework, and make conclusions. | Mature frameworks are missing to audit AI properly. These frameworks need to be generally used, but suitable for specific AI systems. Technical knowledge regarding AI is also lacking. | Depends per client. Start-up AI vendors want to provide evidence but might not have this due to lack of experience. Mature AI vendors don't want to share to much confidential information. | Not enough skills and knowledge. But people are getting trained currently. | There is no one big framework, but many different ones (IIA, KPMG). It is hard to have one uniform framework for all AI systems. Organizations adopt one and adjust it to their needs. |
| Expert 3 | Partner ITAA | 15 year | AI audits are currently performed, both within DSA/DMA audits, as in financial statement audits where AI is used in the financial scope. | DMA is about auditng AI profiling techniques of the client, DSA is about auditing AI recommender systems of the client, and within financial statement, it is about auditing bots that are for reports. | Every client uses a very different sort of AI, which requires us to personalize the audit to this AI, but there are no frameworks for every specific AI system. | No, clients often don't have the overview of all their algoritmes and not everything is clearly documented. Mostly difficult for the big organizations, but they are working on this. | There is a basis, but IT auditors need training, especially on use cases. This is needed, because AI auditing will grow in the future. | There are frameworks which work fine, but you need to tailor them to the context. Controls regarding governance, accuracy, and data selection, testing, and training is needed within the framework. |
| Expert 4 | Consultant RAI | 1 year | A few AI audits are done, but mostly risk assessments. | These risk assessments are about AIA compliance and ethics, so whether the AI system performs in a fair way. | There are not enough proper guidelines for auditing or assessing AI systems. Also, the average IT auditor has not enough knowledge about AI to audit or assess it. | Depends per client, public sector is happy to share evidence, but corporate clients not. It is hard to actually use the code to audit on AIA provisions. | No, the average IT auditor does not have this knowledge. But it is hard to actually audit on aspects like fairness. When is AI fair or not, I find that hard myself, and I studied AI. | IAMA and FRIA exist for AI impact assessments. But guidelines for AI auditing on the AIA are lacking practical guidance, so how to actually perform the audit. |
| Expert 5 | Director ITAA | 12,5 year | Audited AI within DSA/DMA audits, and it will grow due to the demand for AI agents within the current audit and risk workflow, also within financial sector. SAAF project develops this. | AI audit itself is quite new. But within DSA/DMA audits, recommender systems of social media platforms are audited. They look at the parameters of this system for example. | Due to rapid AI developments, more technical understanding of AI is needed within ITAA. More mature frameworks are needed for AI auditing. Auditability is a challenge, and hard to provide AI assurance on the future, when auditing now. | AI vendors are not used to being audited, so they don't document everything good enough, but they also want to spend their time on other things. But this will improve, also due to DSA/DMA audits. | In the future, we will need more auditors with a special profilet to be able to audit AI. Our education program needs to be adjusted to this. | KPMG frameworks are fine, but it is generic. NOREA needs to publish specific AI frameworks. ForHumanity has a certification scheme for AI audits. |

Figure A.1: The first part of the thematic coding.

| Interviewees | Job title | Experience | Frameworks characteristics | HLIC audit | Demand | Risks | Future changes |
|---|---|---|---|---|---|---|---|
| Expert 1 | Consultant RAI | 1 year | It is hard to interpretate the legal texts of the many provisions and translate these to audit controls. A combi of qualitative and quantitative testing is needed, and that is what we are currently working on. | Focus on auditing on bias and fairness. | Mostly governmental organizations will want to audit their AI systems, because they want to create trust within society. I think, if there's no audit obligation, private companies won't do an AI audit on the AI Act, because there's no incentive. | Organizations won't manage their AI risks properly, which may lead to biased or discriminatory outcomes, especially within high-risk AI. Companies could also risk fines by the EU and reputational damage. | The EU should obligate organizations to audit their AI systems. |
| Expert 2 | Manager RAI | 7 year | Frameworks are quite fragmented. But every situation is different, so not every question or risk is relevant for a client. It is doable to implement the AIA provisions into an AI audit framework, but it depends on to what extent the client wants to be risk assured. | Look from the view of the market surveillance authority, AP, who considers it more serious when AI discriminates than when it violates transparency rules. | I think the demand will increase, as I see how internal auditors ask us more for AI audit help. There is no audit obligation, but now, critical sector organizations are liable for their AI, and if they let a third-party audit this independently, they would be more assured and seen as trustworthy. | AI systems could not be compliant to the AIA, and they could maybe make discriminatory decisions that could have a big impact on people's daily lives. | More standards such as the ISO 421001 and from the EU must be publsihed, to provide guidance in how to be compliant to the AIA. |
| Expert 3 | Partner ITAA | 15 year | These frameworks can be used. It is mostly about auditing the processes around the AI system's use, and how we can check that the client is in control over its systems. Concrete examples are missing in the framework, this would help auditors with assessments. | Make sure that the own risk assessment of the client is true; is the system really high-risk? Also, check if training data is valid and there is no data/ model drift. | Demand will increase. Organizations will receive questions from clients about their AI. So if they show an assurance report of an independent audit, this will create trust. | This could result in consequences for society as for organizations. Non-compliance will result in penalties and this leads to a decrease in trust of clients. | The EU must be more clear how and who has to perform the conformity assessment for providers of AI systems. |
| Expert 4 | Consultant RAI | 1 year | IT auditors are used to do interviews and check controls, but for an AI audit framework, they will need quantative testing for data representativeness and fairness. | Aspects like data representativeness, fairness and correct use is important ot audit on. | Demand will slowly increase, but depends on when the market surveillance authority will have the power to check AIA compliance. Until then, it will be mostly audits for public sector organizations that want to avoid reputation damage. | AI systems used in critical sectors could make biased predictions, which could have a huge impact on society and create reputational damage. | AI trainings for KPMG employees would be good, especially when AI auditing demand will increase. |
| Expert 5 | Director ITAA | 12,5 year | These specific AI frameworks must be usable for specific organizations or sectors, because the context and intended use of an AI system determines which risks and controls are relevant. | Currently, AI governance model are introduced within client's organization, where AIA requirements are covered. Also, new AI goes to a risk classification funnel. And, yearly assessments are done. This can be implemented for AIA audits. | Clients could also obligated to audit their AI within DSA/DMA or financial audit, and then, the board might consider to let us check their AI on the AIA, since they are liable for their AI. | If things go wrong due to AI use, it could have an impact on a lot of people. And, reputional damage for organizations is a big risk. | An AI risk&control framework is needed that can be used by multi-internationals worldwide, due to different laws worldwide. Maybe AI audit needs obligations, but you also want to create awareness to do this out of intrinsic motivation. |

Figure A.2: The second part of the thematic coding.

# Appendix B

# Synthetic data

In this section, we present the technical documentation of the NexaSure case study.

## Technical documentation

1. Purpose, provider, and version

The system, named RiskQuant v3.2, is an AI-driven risk assessment model developed by NexaSure, a health and life insurance company. The purpose of the system is to assess the risk profile of insurance applicants and support underwriting decisions by generating a risk score and related explanatory outputs.

2. Interaction with other software and hardware

RiskQuant v3.2 interacts with multiple internal and external systems, including:

- Customer intake systems (application forms, digital onboarding platforms)
- Data integration layers (APIs fetching medical records)
- Underwriting systems (decision support tools used by underwriters)
- Logging and monitoring systems (for compliance and auditability)

The system runs on a cloud-based infrastructure and communicates via secure APIs with back-end databases, analytics engines, and visualization dashboards.

3. Software/firmware versions and update requirements

The current version of the system is v3.2. Previous versions of the system were v3.1 and v3.0, which are archived and retrievable for audit purposes. The updates include the following:

- Model retraining (performed quarterly)
- Security patches (performed monthly)
- Feature updates (performed bi-annually)

4. Forms of system deployment

RiskQuant v3.2 is offered in multiple deployment formats:

- Embedded system: Integrated directly into NexaSure's underwriting software platform.
- API service: Accessible through REST APIs for internal applications and partner systems.
- Web based interface: Used by underwriters to review risk assessments. The user-interface of this is shown in figure 1.

Figure B.1: Page 1 of the technical documentation.

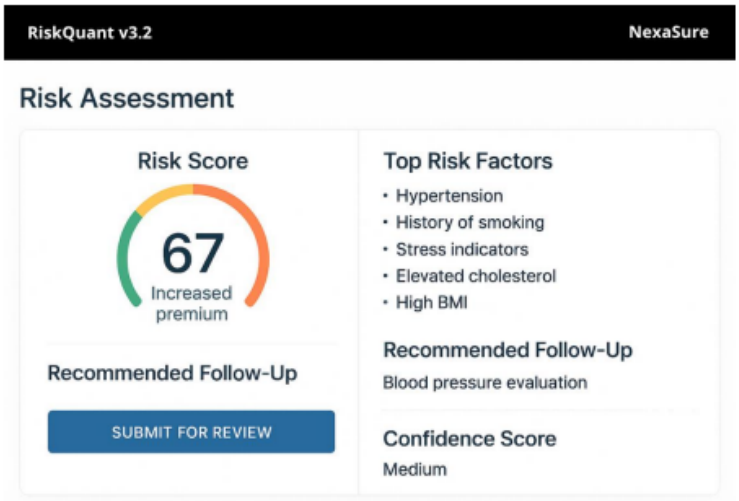



Figure 1. The user-interface of RiskQuant v3.2 when the risk assessment is done.

5. Hardware requirements

The system is deployed on cloud-based infrastructure, typically consisting of:

- Cloud-based servers (Linux virtual machines)
- GPU-enabled environments (Azure Databricks)
- Secure storage system (Archil)

Underwriters access the system via a standard desktop with secure authentication.

6. User-interface description

The system provides a dashboard as the user-interface for underwriters, including:

- A risk score visualization (from 0–100)
- A decision classification (low, medium or high premium, or manual review)
- A list of key risk drivers (explainability output)
- Recommended follow-up actions
- Confidence score of the model (low, medium, or high)

The UI is designed to support interpretability, enable human oversight, and ensure compliance with transparency requirements.

7. Instructions for use

The underwriter must follow these steps:

Figure B.2: Page 2 of the technical documentation.

- Validate the input data of the applicant. Ensure all applicant data is complete and accurate before submission.
- Submit the application through the system to start the assessment.
- Review the outputs of the model and evaluate the risk score, risk drivers, and confidence level.

Then, human oversight processes must be applied. For high-risk cases, which are applicants with a score of 81–100, a mandatory manual review will be conducted. Here, the underwriter assesses whether the AI output aligns with its professional judgment. After this, the underwriter decides whether to accept the applicant for a high premium, accept the applicant for a lower premium, or decline the applicant (if applicable).

This decision must be documented by the underwriter. The underwriter documents the rationale for its decision, especially when overriding AI recommendations. Lastly, the underwriter must report anomalies or unexpected outcomes and provide feedback for model improvement.

8. Procedures and actions taken to build the system

The development of RiskQuant v3.2 followed a structured AI lifecycle approach, including these steps:

- Data collection and preparation from internal and third-party sources.
- Feature engineering to derive relevant medical and financial indicators.
- Model selection and training using supervised machine learning techniques.
- Validation and testing, including accuracy, fairness, cybersecurity, and robustness checks.
- Deployment in a controlled environment with staged rollout procedures.
- Continuous monitoring and improvement, incorporating feedback from underwriters.

All stages are documented and aligned with regulatory requirements.

9. System design, logic, and trade-offs

RiskQuant v3.2 uses gradient boosted decision trees for tabular structured data. Besides this, the model uses rule-based constraints to enforce actuarial and regulatory boundaries.

The design trade-offs for compliance are as follows:

- For the accuracy vs. explainability trade-off, tree-based models are preferred for interpretability.

Figure B.3: Page 3 of the technical documentation.

- For the automation vs. human oversight trade-off, high-risk outputs require mandatory human review.
- For the performance vs. fairness trade-off, bias mitigation techniques are applied where necessary.
- For the complexity vs. auditability trade-off, model transparency is prioritized to enable auditing.

The overall goal is to produce reliable, interpretable, and compliant risk assessments regarding insurance contracts of applicants.

10. System architecture and computational resources

The system follows a modular architecture, which consists of:

- A data ingestion layer
- A feature processing pipeline
- A model inference engine
- A decision logic layer consisting of business rules
- A user-interface
- Monitoring and logging systems

The system runs on the following computational resources:

- Cloud-based servers (Linux virtual machines)
- GPU-enabled environment (Azure Databricks)
- Scalable storage system (Archil)

11. Data and training methods

The system is trained on a combination of the following data, if consent is provided:

- Personal data
- Historical underwriting and claims data
- Medical and health-related data

As for training methods, supervised learning with labeled outcomes is used to train the model. Cross-validation techniques are used to measure the performance of the model. This performance is improved with hyperparameter optimization. The model is retrained on new available data every quarter.

Data is evaluated for:

- Accuracy. This measured with the RMSE and R-squared method, each for different subgroups of the target audience.

Figure B.4: Page 4 of the technical documentation.

- Representativeness. This is measured with the population stability index (PSI).
- Potential bias or imbalance. This is done by comparing the True Positive Rate (TPR) of different subgroups.

| Group | RMSE | R-squared | PSI | TPR |
|---|---|---|---|---|
| Everyone | 0.15 | 0.76 | 0.12 | 0.77 |
| Age 18-39 | 0.18 | 0.83 | 0.16 | 0.84 |
| Age 40-59 | 0.14 | 0.75 | 0.11 | 0.79 |
| Age 60+ | 0.12 | 0.70 | 0.16 | 0.73 |

Table 1. The results of the evaluation of RiskQuant v3.2 of Q3.

12. Technical human oversight measures

RiskQuant v3.2 integrates human oversight through the following elements:

- Mandatory human review for high-risk scores
- Override functionality for underwriters
- Explanation outputs, which are the risk drivers, to support decisions
- Alert mechanisms for anomalies or low-confidence predictions

13. Pre-determined modifications

The system includes predefined mechanisms for updates and changes:

- Scheduled model retraining (performed quarterly)
- Version updates (v3.2 to v3.3)
- Adjustment of decision thresholds (risk score ranges)

All pre-determined modifications are tested before deployment and documented in version control systems.

14. Cybersecurity measures

The system incorporates multiple layers of cybersecurity:

- Data encryption
- Role-based access control (RBAC)
- Multi-factor authentication (MFA)
- Secure API communication protocol Model Context Protocol (MCP)
- Intrusion detection and monitoring systems
- Regular penetration testing and vulnerability assessments

15. Unintended outcomes and risks

Figure B.5: Page 5 of the technical documentation.

There are a few potential unintended outcomes and risks to health, safety, and fundamental rights of applicants that could arise due to the use of RiskQuant v3.2. These are as follows:

- There is a risk that certain subgroups significantly receive higher risk scores than other subgroups. This is a risk to the fundamental rights of individuals if the subgroups are based on characteristics such as location, gender, and nationality.
- There is a risk that the previously mentioned bias provides new biased training data for the model when it is retrained.
- There is a change that certain subgroups, discriminated against by RiskQuant v3.2, will avoid purchasing insurance and getting medical help, which results in a health and safety risk.
- There is a risk of automation bias. This means that the underwriter relies too heavily on the model's recommendations. This could result in discriminatory outcomes for certain subgroups.
- There is a risk of a data breach within the AI system. If this happens, sensitive personal data could fall into the hands of criminals.

16. RMS

NexaSure has implemented an RMS, from which the user-interface is shown in figure 1. This image shows that NexaSure has a clear overview of its risks, risk mitigation activities and controls, regarding the risks that arise due to the use of RiskQuant v3.2. Besides this, NexaSure has an overview of the roles and responsibilities within the company, regarding these risks. This overview is illustrated with the RASCI Matrix.

Figure B.6: Page 6 of the technical documentation.

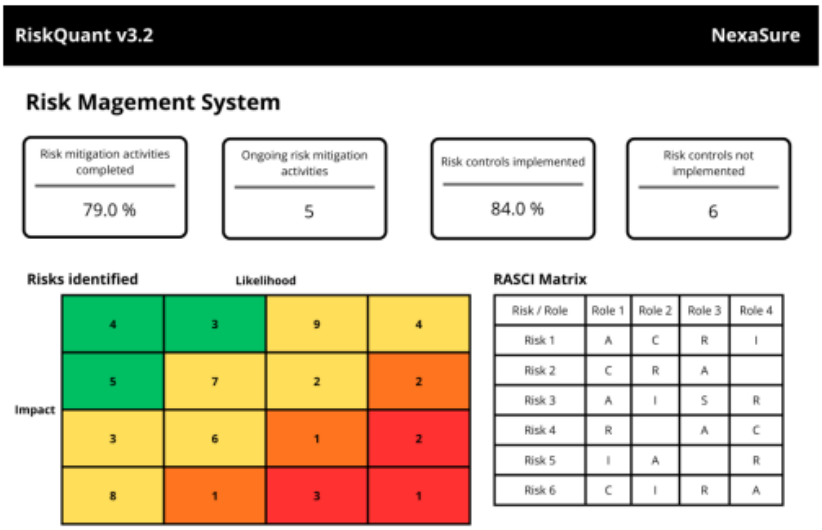



Figure 2. The user-interface of the RMS.

17. Harmonized standards or solutions

Since the harmonized standards for the AI Act are not yet published by the CEN and CENELEC, another framework is used to ensure compliance with the AI Act. The AI auditing framework of the Institute of Internal Auditors (IIA) was used for the support of the process.

18. Post-Market Monitoring Plan

Post-Market Monitoring System (PMMS) is implemented by NexaSure. This system is able to monitor the functionality of the system and is proportionate to the risks regarding health and life insurance. Every malfunction that occurred or risk that has been realized is registered here, which helps Nexasure manage and resolve these disruptions. Deployers of the risk assessment model provide system data to the PMMS, which supports NexaSure in being compliant with the AIA continuously. This data includes the R2 and ROC-AUC score of the risk assessment model. This is measured per quarter, since the model is retrained

Figure B.7: Page 7 of the technical documentation.

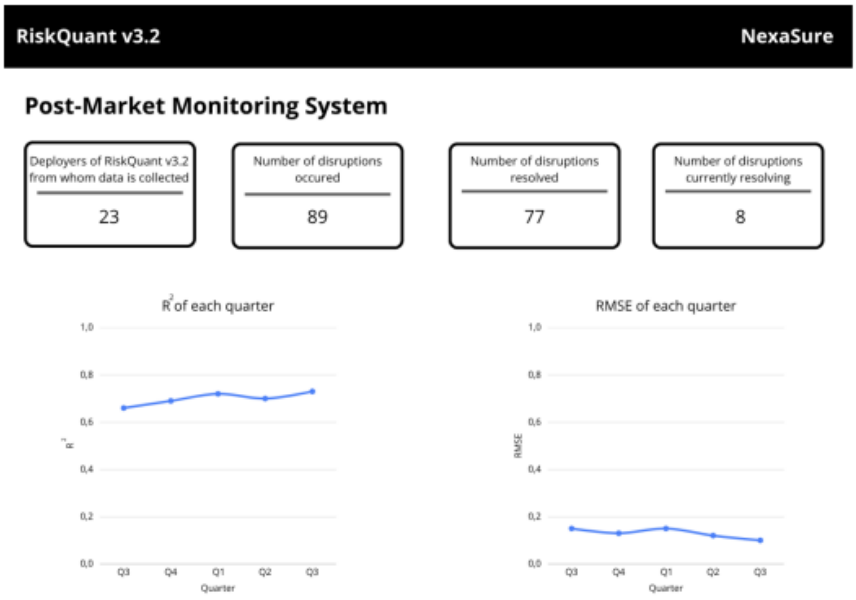



Figure 3. The user-interface of the Post-Market Monitoring System (PMMS).

Figure B.8: Page 8 of the technical documentation.

# Appendix C

# Audit controls

Here, the screenshots of the Audit controls component of the AI audit framework are shown.

| Control reference | Control name | General control description | Article reference | Control procedure description |
|---|---|---|---|---|
| 1.1 | RMS general | Auditor must assess the implemention of the Risk Management System (RMS). | 9.1, 9.2 | Check if the RMS is able to identify, assess, and mitigate risks, including risks that have an impact on the health, safety and fundamental rights of individuals. |
| 1.2 | RMS characteristiscs | Auditor must assess the RMS's characteristics. | 9.5, 9.6, 9.7, 9.8, 9.9 | Check if the the actions of the RMS have these requirements:<br>- The goal of the action is to remove of minimize the risk.<br>- The actions include mitigations for risk that cannot be removed.<br>- The actions include information and training for deployers, considering their knowledge and experience.<br>- The residual risk is acceptable.<br>- The RMS considers potential negative impact on vulnerable users.<br>- The system is frequently tested on its RMS, to ensure compliance with Article 9. |
| 2.1 | Data governance | Auditor must assess the training, testing and validating datasets. | 10.2 | Check if the three datasets are managed by data governance and management procedures, including these elements:<br>- Clear design choices in relation to the purpose of risk assessment within health or life insurance.<br>- The processes of data collection, including the origin of the data. The purpose of data collection is needed if personal data is involved.<br>- The assessment of the datasets quantity, suitability, and availability.<br>- The measures for bias identification and mitigation.<br>- The assessment of potential biases that could harm the health, safety, and fundamental rights of individuals.<br>- The process of identifying any missing or insufficient data that would hinder compliance with the AIA, including the potential measures for resolving these gaps. |
| 2.2 | Data quality | Auditor must assess the representativeness of the traning, testing and validating datasets. | 10.3 | Check if the three datasets meet the following requirements for as much as possible:<br>- The data is accurate and complete for the risk assessment within health or life insurance.<br>- The datasets possess proper statistical properties (sample size, distribution, variance), including a representative reflection of the different subgroups of applicants. |

Figure C.1: The first columns of the Audit controls part 1.

| Needed audit evidence | Control guidance | Conclusion | Control gaps | Guidance AI |
|---|---|---|---|---|
| - An overview of the RMS, including the processes and capabilities.<br>- The AI governance framework.<br>- The roles and responsibilities regarding AI risks. | | Positive | | Possible risks when using AI systems are:<br>- Models that proces data without consent<br>- Biased or discriminatory models<br>- Lack of transparancy of the model (Black box)<br>- Models that produce wrong output (Hallucinations) |
| Documentation, screenshots and walkthroughs regarding the elements of the RMS, including its processess, goal, capabilities, residual risk, and testing. | | Positive with comments | | |
| Samples from relevant datasets and information regarding design choices, processes, and bias detection procedures of these datasets, including metrics and results of bias detection. | - This only applies for testing datasets if the provider or deployer does not develop the system using training techniques.<br>- Use the Aequitas tool to assess fairness within the model. This can be done by calculating the disparate impact ratio for example. | Negative - except for | | Training data is used to build the model, testing data is used to evaluate performance, and validation data is used to fine-tune and prevent overfitting. Bias means that the model produces systematically unfair and inaccurate results, disadvantaging certain subgroups. This is unfair when disadavantages are based on data that is not allowed for training |
| - Information about the target population of provider.<br>- Data samples from these datasets.<br>- Information regarding test splits.<br>- Statistical properties of the datasets.<br>- Documentation regarding data completeness. | - This only applies for testing datasets if the provider or deployer does not develop the system using training techniques.<br>- Calculate proportion per subgroup in dataset and compare this with real applicant distribution, to check if the data is accurate.<br>- Calculate percentages of missing data per subgroup, these cannot differ significantly, to check data completeness and representativeness.<br>- Calculate the PSI (Population Stability Index) to check if data drift occured in the data, to check if data drift occured. | Negative - adverse | | |

Figure C.2: The last columns of the Audit controls part 1, with the four possible scoring labels as an example.

| | | | | |
|---|---|---|---|---|
| 2.3 | Data selection | Auditor must assess the data selection of the training, testing and validating datasets. | 10.4 | Check if the three datasets only contain data that reflects the environment of risk assessment within health and life insurance. This can include age, health, medical, or actuarial claim data, but no marketing, sales, social, or behavioural data. |
| 2.4 | Bias detection | Auditor must assess the bias detection procedure. | 10.5 | Check if the provider processes special categories of personal data, if this is needed for bias detection and correction, under the following conditions:<br>- The bias causes harm to the safety, health of fundamental rights of individuals.<br>- The bias cannot be detected and corrected by processing other data.<br>- The re-use of special categories of personal data is technically restricted and safeguarded through security and privacy-protecting measures.<br>- The special categories of personal data will not be shared with other parties.<br>- The special categories of data will be deleted when its retention period is reached or the bias is corrected.<br>- The reasons why the processing of special categories of personal data is needed to detect and correct bias. |
| 3.1 | Technical documentation general | Auditor must determine if sufficient information of the system is provided in the technical documentation. | 11.1, Annex IV (1) | Check if the technical documentation updates continuously and includes the following:<br>- The purpose of the system and the name of the provider.<br>- How the system interacts with other software and hardware.<br>- Details of the software or firmware versions and any associated update requirements.<br>- An overview of every form in which the system is offered or used.<br>- A general description of the user-interface that is provided to the deployer of the system.<br>- Instructions for the use of the system for the deployer. |
| 3.2 | Precise technical documentation | Auditor must assess the detailed descriptions in the technical documentation, regarding the development of the system. | 11.1, Annex IV (2, 8) | Check if the technical documentation includes precise descriptions of the following:<br>- The procedures and actions taken to build the system.<br>- The specifications of the system's design, logic, classification strategies, optimization goals, and the reasoning behind any trade-offs made for compliance.<br>- An overview of all the methods used for training.<br>- An overview of the technical human oversight measures.<br>- An overview of the validation and testing processes.<br>- All potential unintended outcomes and risks to health, safety, and fundamental rights, and the used metrics to identify these.<br>- An EU declaration of conformity. |

Figure C.3: The first columns of the Audit controls part 2.

| | | | | |
|---|---|---|---|---|
| Samples from the three datasets. | - This only applies for testing datasets if the provider or deployer does not develop the system using training techniques. | | | |
| - Information regarding the bias and all the required reasons why the processing special categories of personal data is needed.<br>- Evidence regarding the confidentiallity and removal of the personal data is needed. | | | | |
| A technical documentation file or screenshots of this, including all the needed information of the system, updates, hardware, user-interface, and user instructions. | | | | AI runs on software and hardware. The software loads the model, sends data into it, receives results and uses these. The hardware provides the computing power for the calculations. |
| Detailed descriptions of all the elements that are listed in the cell left from this one. | | | | Actions taken to build AI are data collection and preparation, model selection, training process, and testing and validation. The design means the structure of the model, and the logic is about the workflow of how inputs lead to outputs. Classification strategies tells us how predictions are made exactly, and optimization goals tells us what aspect the model tries to maximize (e.g. accuracy, fairness, speed or explainability. Training methods can be supervised or unsupervised and are classification or regression methods. Human oversight measures can be manual review of cases and override mechanisms. |

Figure C.4: The last columns of the Audit controls part 2.

| | | | | |
|---|---|---|---|---|
| 3.3 | Performance metrics documentation | Auditor must assess the performance metrics within the technical documentation. | 11.1, Annex IV (4) | Check if the performance metrics regarding accuracy, robustness and compliance of the system is appropriate and included in the technical documentation. |
| 3.4 | RMS documentation | Auditor must assess the RMS description within the technical documentation. | 11.1, Annex IV (5) | Check if the RMS is described in detail within the technical documentation. |
| 3.5 | Harmonized standards documentation | Auditor must assess the used harmonized standards for compliance within the technical documentation. | 11.1, Annex IV (7) | Check if all the used harmonized standards that are published in the Official Journal of the European Union are descripted in the technical documentation. If no harmonized standard is used, any other solution must be provided for compliance with the AIA. |
| 4.1 | Record-keeping general | Auditor must asses the record-keeping processes of the system. | 12.1 | Check if the system is able to record events automatically and during the entire lifecycle of the system. |
| 4.2 | Events record-keeping | Auditor must assess the capabilities of the record-keeping of the system. | 12.2 | Check if the logging processes of the system enable the recording of actions relevant for:<br>- Detecting situations that could lead to a substantial change or a risk regarding the health, safety or fundamental rights of the applicant caused by the system.<br>- Facilitating the post-market monitoring process.<br>- Monitoring the performance of the high-risk system. |
| 5.1 | Transparency general | Auditor must asses the transparancy of the system. | 13.1 | Check if the system is designed and developed so that its functioning is transparant enough for you as auditor to understand the system's output and use them correctly. |
| 5.2 | Instructions for deployer | Auditor must assess the user instructions that are provided by the system. | 13.2 | Check if the user instructions are provided by the system, in a proper digital or other format. These instructions must be precise, accurate, and easy for non-AI-literate deployers to access and understand. |

Figure C.5: The first columns of the Audit controls part 3.

| | | | | |
|---|---|---|---|---|
| A description of the performance metrics of the system within the technical documentation. | Use the R-squared and RMSE method to measure the performance of the risk assessment model. | | | Metrics for classification models are Accuracy and F1-Score. Metrics to use for regression models are R-squared and RMSE. Calculate per subgroup to measure bias within the training data. |
| A description of the RMS within in technical documentation. | | | | |
| A overview of the used harmonized standards or solutions for compliance with the AIA, within the technical documentation. | | | | |
| An overview of the record-keeping processes, including the logs. | | | | The AI system can record events such as:<br>- Modification of parameters<br>- Input data received<br>- Model prediction process<br>- Model output generation<br>- Decision of the output |
| Information regarding the record-keeping proces, including its capabilities and limitations. | | | | |
| Information about the functioning and the outputs of the system, including all the information that is provided to the user. | | | | To determine whether an AI system is transparent, you assess whether non-AI-literate persons can understand how it makes decisions and how these are used. Besides this, the purpose of the system should clear, its processes are documented, and its decisions can be explained. |
| Screenshots of the user instructions that are provided by the system. | | | | |

Figure C.6: The last columns of the Audit controls part 3.

| | | | | |
|---|---|---|---|---|
| 5.3 | Instruction elements | Auditor must assess the elements of the user instructions that are provided by the system. | 13.3 | Check if the user instructions include the following elements:<br>- Details about the provider, including contact information.<br>- The intended purpose of the system.<br>- The tested and validated levels of accuracy, including the metrics used, as well as the levels of robustness and cybersecurity.<br>- Any known or foreseeable situation that could lead to risks regarding the health, safety, or fundamental rights of the user.<br>- Where applicable, technical features and information that help the deployer to understand and explain the system's outputs.<br>- Details regarding system's input data.<br>- Any pre-defined modifications to the system at the time of the original conformity assessment.<br>- The human oversight measures that are required by Article 14.<br>- The required computational and hardware sources, the expected operational lifespan of the system, and the needed maintenance activities for the functionality.<br>- Where applicable, an explanation of the features built into the system that enable deployers to correctly gather, save, and understand the logs in line with Article 12. |
| 6.1 | Human oversight general | Auditor must assess the overall human oversight process of the system. | 14.1, 14.2 | Check if the system is created so that individuals can effectively supervise them while deploying. This should include proper human–machine interface tools for support. The goal of this process should be to prevent or reduce risks to health, safety, or fundamental rights that may arise when the system is used. |
| 6.2 | Human oversight measures | Auditor must assess the human oversight measures of the system. | 14.3 | Check if the human oversight measures match the level of risk and autonomy of the system, within the context of risk assessment within health or life insurance. The measures should be one or both of the following types of measures:<br>- Measures that the provider detects and implements before the system is placed on the market.<br>- Measures that the provider detects after the system is placed on the market and that are suitable to be implemented by the deployer. |
| 6.3 | Human oversight capabilities | Auditor must assess the human oversight capabilities of the system. | 14.4 | Check if the system is delivered to the deployer in a way that allows an individual that is responsible for human oversight to perform the following actions:<br>- Understanding the output, relevant abilities and weaknesses of the system, and effectively monitoring its behavior.<br>- Staying conscious of automation bias.<br>- Deciding to ignore the outcomes of the system or to stop working with the system.<br>- Stepping in and taking control of the high-risk AI system's functioning, or stopping it using a 'stop' button or a similar method that allows the system to safely shut down. |

Figure C.7: The first columns of the Audit controls part 4.

| | | | | |
|---|---|---|---|---|
| Screenshots and information regarding all the elements of the user instructions that are provided by the system. | | | | |
| Screenshots and information regarding the human oversight process, including the human-machine interface tools. | | | | |
| Screenshots and information regarding the human oversight measures, including the type of measure. | | | | |
| Screenshots and information regarding the capabilities of individuals that are responsible for human oversight, and the information that is provided to them. | | | | Automation bias within the use of AI means the tendency to trust and use AI outputs without questioning them, even when they may be wrong. |

Figure C.8: The last columns of the Audit controls part 4.

| | | | | | |
|---|---|---|---|---|---|
| 7.1 | Accuracy | Auditor must assess the accuracy of the system. | 15.1, 15.2, 15.3 | Check if the accuracy of the model is appropriate and performs consistently. Obligated levels and methods for the accuracy will be published by the EU in the future, so until then, this matter is left to professional judgment. | The used metrics and samples of the model to calculate its accuracy. |
| 7.2 | Robustness | Auditor must assess the robustness of the system. | 15.1, 15.4 | Check if the system is built to reduce errors, malfunctions, and inconsistencies that may arise, for as much as possible. Check if organisations have implemented both technical and organisational safeguards. The design must ensure that biased outputs do not influence future inputs in a way that creates harmful feedback loops. Any feedback loop that could still occur must be identified and controlled through proper mitigation solutions. | An overview of all the robustness measures that are implemented in the system. This include code checks for measures that reduce erros, malfunctions, and inconsistencies. |
| 7.3 | Cybersecurity | Auditor must assess the cybersecurity measures of the system. | 15.5 | Check if the system is capable of withstanding attempts by unauthorized third parties to exploit vulnerabilities in order to change the model's processes or outcomes. The cybersecurity measures put in place must be suitable for risk assessment within health and life insurance and the risks that are involved in this field. The technical measures should include mechanisms, where relevant, to avoid, identify, react to, mitigate, and control attacks. | An overiew of all the cybersecurity measures that are implemented in the system. |
| 8.1 | QMS | Auditor must assess the Quality Management System (QMS) of the system. | 17.1, 17.2, 17.3, 17.4 | Check if a QMS is implemented and maintained in the system. The QMS must be thoroughly documented in an organized and structured way, using written policies, procedures, and instructions. This should include aspects such as development, testing and post-market monitoring activities. | Information and screenshots of the QMS, including the user-interface, documentation, processes and capabilities. |
| 9.1 | Corrective actions | The auditor must determine whether the provider appropriately handles corrective actions. | 20.1 | Check if the provider takes the necessary actions when its system is not compliant to the AIA or a risk is presenting. | An overview of all occurrences of non-compliance with the AIA or identified risks, together with the corrective actions taken by the provider. |
| 9.2 | Duty information | The auditor must determine whether the provider appropriately fulfills its duty of information. | 20.2 | Check if the provider informs deployers, distributors, and authorized representatives, when the system is not compliant to the AIA or a risk is presenting. | An overview of all communications with deployers, distributors, and authorized representatives at the time of non-compliance with the AIA or risk occurrences. |

Figure C.9: The first columns of the Audit controls part 5.

| | | | | |
|---|---|---|---|---|
| The used metrics and samples of the model to calculate its accuracy. | | | | |
| An overview of all the robustness measures that are implemented in the system. This include code checks for measures that reduce erros, malfunctions, and inconsistencies. | | | | |
| An overiew of all the cybersecurity measures that are implemented in the system. | | | | |
| Information and screenshots of the QMS, including the user-interface, documentation, processes and capabilities. | | | | |
| An overview of all occurrences of non-compliance with the AIA or identified risks, together with the corrective actions taken by the provider. | | | | |
| An overview of all communications with deployers, distributors, and authorized representatives at the time of non-compliance with the AIA or risk occurrences. | | | | |

Figure C.10: The last columns of the Audit controls part 5.

| | | | | |
|---|---|---|---|---|
| 10.1 | Supervisory authorities cooperation | The auditor must assess the cooperation of the provider with supervisory authorities. | 21.1, 21.2, 21.3 | Check if the provider fully cooperated with its supervisory authority, when this was requested by the supervisory authority. |
| 10.2 | Authorized representative | If the provider is established outside the EU, the auditor must determine whether the provider appointed an authorized representative. | 22.1, 22.2 | Check if the provider, if established outside the EU, appointed an authorized representative by a written mandate. This authorized representative must be located within the EU and must be enabled by the provider to perform the tasks written in the mandate. |
| 11.1 | PMMS | The auditor must assess the Post-Market Monitoring System (PMMS) of the system. | 72.1, 72.2 | Check if the provider implemented a PMMS with the following requirements:<br>- It is proportionate to the size and the risks of the system.<br>- It is based on the post-market monitoring plan, which is part of the technical documentation.<br>- It is able to document and analyze relevant data, that is provided by the deployer, about the performance of the system. |
| 12.1 | Conformity assessment | The auditor must determine whether the provider conducted the conformity assessment correctly. | 43.2, Annex VI | Check if the provider has verified that the system meets the requirements of Article 9-15, 17, and 72. |
| 13.1 | Reporting incidents | The auditor must determine whether all serious incidents are reported by the provider. | 73.1 | Check if the provider reported every occured, serious incident to the market surveillance authority of the country that it occurred in. A serious incident refers to any event or failure of a system that, either directly or indirectly, results in one or more of the following outcomes:<br>- The death, or significant damage to the health, of an individual.<br>- A major and irreversible disruption to the functioning or management of critical infrastructure.<br>- A breach of provisions under EU law that are designed to safeguard fundamental rights. |

Figure C.11: The first columns of the Audit controls part 6.

| | | | | |
|---|---|---|---|---|
| An overview of all communications with the supervisory authority of the provider. | | | | |
| All the documents regarding the assigning of the authorized representative. | | | | |
| Screenshots and information regarding the PMMS and the post-market monitoring plan, including its processes and capabilities. | | | | |
| All the documentation that shows the provider verified compliance with Article 9-15, 17, and 72. | | | | |
| An overview of all communications with the relevant market surveillance authority, when a serious incident occurred. | | | | |

Figure C.12: The last columns of the Audit controls part 6.