From Principles to Practice:

A Tailored Maturity Model for Responsible AI in Public Sector Organisations

2025

Graduation Committee:

UT - Robin Effing

UT - Faiza Buksh

PwC - Amine Achahboun

Author:

Melvin Willems

PwC | EEMCS & BMS

Acknowledgements

Completing this Master's thesis has been an incredibly informative journey, and I am happy to have chosen Responsible AI as the focus of my research. It is a fascinating field, intersecting with disciplines such as philosophy, computer science, public administration, and information systems. The relevance of this topic will only continue to grow in the coming years, and I hope the findings of this thesis contribute to both academic and practical advancements in understanding Responsible AI.

This achievement would not have been possible without the support and guidance of many individuals. I am grateful to everyone who dedicated their time and efforts to participate in the Delphi study and the subsequent validation study.

I would like to extend my gratitude to my university supervisors, Robin Effing and Faiza Buksh, for their guidance, encouragement, and constructive feedback. Beyond the academic aspects, I really appreciated our personal conversations and the supportive environment you created throughout this process.

I am also super thankful to PwC for providing me the opportunity to complete my thesis within their organisation. A special thanks to Amine Achahboun and Tosja Selbach for sharing your expertise, offering your support, and involving me in client projects and discussions.

On a personal note, I am incredibly thankful to my family and friends for their encouragement and support throughout this journey. A special mention goes to Agata – your support and patience with me have meant the world to me.

Lastly, a brief acknowledgment to Lourenço (2021) for developing the LaTeX template that was used for this thesis.

Abstract

Purpose: This thesis introduces the Responsible AI Maturity Model (RAI-MM), specifically designed for public sector organisations to bridge the gap between high-level ethical principles and their practical implementation. The rapid advancements in AI, growing labour shortages, and the emergence of new AI regulations underscore the need for new tools.

Design/Methodology: The research employs a Design Science Research Methodology (DSRM), following the procedural model outlined by Becker et al. (2009). The study unfolds in three key phases: (1) a systematic review and comparison of 22 existing AI maturity models, (2) the creation of a novel maturity model through a three-round Delphi study involving experts from academia, consultancy, and public sector organisations, and (3) an empirical validation of the RAI-MM through two case studies and two expert sessions.

Findings: The resulting RAI-MM is an empirically validated framework comprising five dimensions: Strategy, Culture & Competences, Governance & Processes, Data & Information, and Technology & Tooling. These dimensions encompass twenty specific items designed to evaluate and enhance the maturity of public organisations in adopting responsible AI practices. The evaluations demonstrate the adequacy of the model to evaluate responsible AI capabilities with public sector organisations.

Conclusion: The RAI-MM is a practical tool for public sector organisations to assess their responsible AI maturity, facilitate discussions for improvement, and support compliance with the AI Act. This regulatory framework remains challenging for many organisations to implement.

Keywords: Responsible AI, Maturity Model, Public sector organisations

Contents

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence *(p. 8)*

CMM Capability Maturity Model *(pp. 6, 7)*

DSA Digital Services Act *(p. 53)*
DSR Design Science Research *(p. 16)*
DSRM Design Science Research Methodology *(pp. 4, 16, 17)*

GANs General Adverserial Networks *(pp. 9, 10, 13)*
GDPR General Data Protection Regulation *(pp. 47, 53)*
GenAI Generative Artificial Intelligence *(pp. 1, 8–10)*

HCAI Human-Centred Artificial Intelligence *(pp. 13, 34)*
HCI Human-Computer Interaction *(p. 2)*

IS Information Systems *(pp. 2, 7, 14, 16, 17, 21, 22, 24, 78)*

KRNW Knowledge Resource Nomination Worksheet *(pp. 22, 41, 76)*

LLM Large Language Model *(p. 8)*
LLMs Large Language Models *(pp. 12, 13)*

ML Machine Learning *(p. 9)*
MM Maturity Model *(pp. 2, 7, 8)*

NLP Natural Language Processing *(p. 12)*

OECD Organisation for Economic Co-operation and Development *(p. 1)*

QDA Qualitative Data Analysis *(p. 42)*

RNN Recurrent Neural Network *(p. 12)*

SLR Systematic Literature Review *(pp. 19, 27)*

TAM Technology Acceptance Model *(p. 24)*

UTAUT Unified Theory of Acceptance and Use of Technology *(p. 24)*

VAEs Variational Autoencoders *(pp. 10, 11)*

1 Introduction

Ethical discussions often surge in tandem with technological advancements. One illustrative example is the recent debate about Generative Artificial Intelligence (GenAI). This field has sparked numerous ethical dilemmas and stimulated regulators to develop guidelines for Responsible AI. The recent introduction of AI regulations by the European Commission underscores this trend (Parliament & of the European Union, 2024). In this context, organisations, particularly governmental bodies struggle with assessing their ethical awareness and navigating their journey towards ethical maturity (Anagnostou et al., 2022).

Additionally, initiatives like the adoption of intergovernmental policy guidelines on AI by the Organisation for Economic Co-operation and Development (OECD), and the Dutch government's presentation of their vision for GenAI at the beginning of 2024 highlight the global significance of ethical considerations in AI development (Ministry of the Interior and Kingdom Relations, 2024; OECD, 2019). Katz et al. (2023) observed a growing interest in Responsible AI, noting that terms like 'responsibility', 'ethics', and 'bias' frequently appear in client inquiries about AI. They anticipate this interest will further increase as the adoption of AI expands and ethical issues become more prominent.

These concerns have prompted a surge of AI guidelines and codes of ethics in both the public and private sector. However, their holistic and contested nature makes them difficult to apply, resulting in ethics operating at a maximum distance from practice and a gap between high-minded principles and technological practice (Munn, 2022).

1.1 Research problem

Academics have called for a shift from the 'what' of ethics to the 'how' of applied ethics (e.g. Morley et al., 2020; Zhou & Chen, 2022). Despite developing various frameworks and models, their integration into industry practices must be clarified. Morley et al. (2021) concluded that most tools and methods are either too flexible (thus susceptible to ethics washing) or too strict (and unresponsive to context). Qiang et al. (2023) urge the AI ethics community to define existing frameworks' suitability and expected benefits better

to enhance their adoption in industry practices. Although most discussions about AI ethics have occurred outside the Information Systems (IS) field, the topic is closely connected to Human-Computer Interaction (HCI) and organisational culture and processes (Jantunen et al., 2021). The IS community approaches AI ethics through the lens of Responsible AI Vassilakopoulou et al. (2022).

The rapid development of AI, the growing labour shortages (Pouliakas et al., 2024), and the emergence of new AI regulations show the urgency and relevance for new tools, methods, and metrics. In practice, organisations already face challenges related to the responsible development and use of AI. A survey by EenVandaag reveals that 74 percent of civil servants at municipalities in the Netherlands use ChatGPT extensively, despite recommendations against using AI software due to the associated risks (van Wanroij, 2023). Similarly, some organisations have teams developing AI tools, but when these tools are passed on for use in other teams, it is often not clear who carries the responsibility. Although this is a clear requirement outlined in the AI act, organisations are struggling to define this properly.

This research primarily focuses on public organisations as the problem context, due to the limited availability of validated frameworks. The irresponsible use of AI is a significant issue, particularly for governments, as they interact directly with citizens. This leads to the formulation of the following research problem that this thesis aims to tackle:

> *Ensuring the responsible use of AI is crucial as the technology becomes increasingly relevant in the coming years. However, public organisations often need more comprehensive guidelines and frameworks to implement Responsible AI practices effectively. The gap between high-level principles and ethical AI practices creates challenges in adopting and maintaining ethical standards. Therefore, there is a need for a framework that identifies the necessary capabilities an organisation requires, provides guidance for the adoption of Responsible AI, and allows for benchmarking its performance against other public organisations.*

1.2 Research Objective

The method to address the gap described in the problem statement is to provide a tested and validated model that brings ethics closer to technological practices. The focus will be on the public sector, which has expressed a need for guidelines on adopting ethical AI practices. The most suitable method for this research is the development of a Maturity Model (MM). Maturity models are designed to reveal current and desired maturity levels and include respective improvement measures (Pöppelbuß & Röglinger, 2011). Such models intent to diagnose and eliminate deficient capabilities. As Wendler (2012) noted, it is crucial for authors of new maturity models to first assess if any existing models could be relevant. This step is important to improve the quality and relevance of new model developments and prevent unnecessary development expenditure.

At the heart of this research is developing a Responsible AI framework inspired by existing maturity models such as those proposed by Krijger et al. (2022) and Dotan et al. (2024). These models provide a robust foundation upon which to build. However, the existing models have shortcomings that this thesis aims to address. Firstly, neither model is tailored to the public sector, lacking indicators specifically important for this field. An essential aspect for public sector organisations is their interaction with citizens, as it reflects citizen's trust in the government and politics. Additionally, the existing models are not sufficiently detailed. For instance, the model of Krijger et al. (2022) only provides maturity levels for five dimensions without further defined sub-aspects, making it less practical and more arbitrary for organisations to measure their maturity.

Vakkuri et al. (2021) also highlighted the necessity for a Responsible AI maturity model, posing a key design question: Should the model focus on Responsible AI maturity or adopt a broader AI maturity perspective? Each approach has its advantages and disadvantages. Given the significant relevance of technical foundations to ethical adherence, a decision has been made to integrate AI foundations with ethical dimensions. The Responsible AI maturity model will offer a roadmap for public organisations, guiding them from ad hoc implementations to a more mature and standardised approach to their ethical practices and processes.

1.3 Research Questions

The design problem, also known as a technical research problem (Wieringa, 2014), is defined as follows:

> *Improve* Responsible AI practices within public organisations, *by* developing a Responsible AI maturity model *that* translates high-level ethical AI principles into actionable guidelines *in order to* ensure that AI systems provide positive outcomes for citizens without causing negative impacts.

The primary objective of this research is thus to design a maturity model that advocates Responsible AI. The research questions follow the guidelines of Thuan et al. (2019) on constructing design science research questions. The main research question is formulated as follows:

> **RQ:** How can a maturity model be developed and evaluated for public organisations to measure and guide their Responsible AI capabilities?

The question is designed to address both the design and validation of a practical tool for public organisations. The "how" formulation indicates the methodological approach, aligning with design theory, which often aims to produce prescriptive artifacts.

The distinction between "develop" and "evaluate" highlights the dual focus on constructing the model and testing its applicability in real-world scenarios. This ensures the

model is both theoretically sound and empirically validated, setting it apart from existing models. The term "developed" is preferred over "designed" in this design science study to align with the leading methodology of this research, specifically the procedural model proposed by Becker et al. (2009).

The term "guide" is deliberately chosen to indicate that the maturity model provides direction rather than definitive answers, acknowledging the complexity and ongoing discourse in Responsible AI.

In order to design an effective Responsible AI maturity model, the following sub-questions have been formulated:

SRQ 1: What (Responsible) AI maturity models are available in current academic literature?

SRQ 2: What levels, dimensions and items should be included in a Responsible AI maturity model for the public sector?

SRQ 3: How does the Responsible AI maturity model hold up in practice?

The aim of **SRQ 1** is to gain insight into the state-of-the-art Responsible AI maturity models. Answering this research question is crucial for building a cumulative tradition and ensuring that the new maturity model from this research adds value to the academic community. **SRQ 2** focusses on the design and development of the maturity model. It is essential to understand the relevant dimensions and levels to assess how companies can improve their practices, while also tailoring the model to public organisations, which have different needs compared to commercial organisations. **SRQ 3** specifically addresses the creation of a maturity assessment to make it a practical tool for organisations that can be used in a real-world context.

1.4 Research Methodology

This research employs a Design Science Research Methodology (DSRM) approach, further explained in chapter 3. Design Science involves the creation and study of artifacts within a specific context (Wieringa, 2014). These artifacts are designed to interact with a problem context to bring about improvements. Although there are various methodologies for developing maturity models, this research follows the methodology proposed by Becker et al. (2009). Their methodology is specifically aimed at the development of maturity models and is applicable to all types of maturity models, unlike other methodologies that focus either on specific maturity models (e.g. Mettler & Rohner, 2009; van Steenbergen et al., 2010) or are design methodologies for any artifact (e.g. Peffers et al., 2007). The seven steps are visually represented in figure 1.1

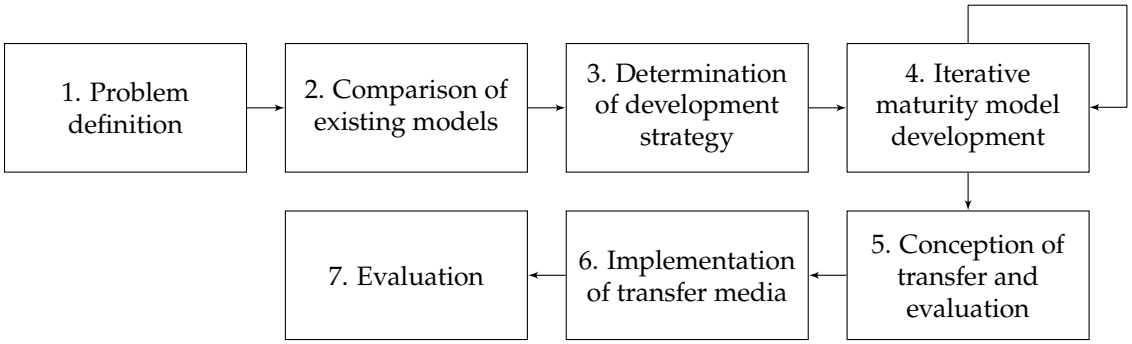


Figure 1.1: Procedural model Becker et al. (2009)

1.5 Research Outline

The remainder of this report is structured according to the procedural steps of Becker et al. (2009):

- **Chapter 2** provides the theoretical foundation of the research. It provides a background in Maturity Models, Artificial Intelligence, and Responsible AI, the relevant disciplines for this thesis.
- **Chapter 3** elaborates on the methodology used to design and evaluate the maturity model.
- **Chapter 4** gives an overview of the existing maturity models in the field of (Responsible) AI, corresponding to the comparison of existing maturity models as defined by Becker et al., 2009. It concludes with a determination of the development strategy and a conceptual maturity model.
- **Chapter 5** describes the iterative development of the maturity model with the Delphi method.
- **Chapter 6** describes the development of a practical tool/assessment for measuring the maturity of organisations.
- **Chapter 7** evaluates and validates the model through various case studies and includes a description of the maturity assessment that was developed after completion of the Delphi study.
- **Chapter 8** discusses the implications of the research.
- **Chapter 9** concludes the report, providing practical and theoretical contributions, as well as the limitations of the research.

2

Theoretical Background

This chapter covers the academic foundation of this research. Section 2.1 offers an overview of maturity models, providing insights into the various types and their objectives. Sections 2.2 and 2.3 focus on Artificial Intelligence and Responsible AI, respectively. Understanding the concept of AI is necessary for defining the boundaries of the field, while understanding Responsible AI is essential for comprehending the ethical principles and research paradigms associated with it.

2.1 Maturity Models

Many research papers refrain from clearly defining maturity models (Wendler, 2012). Instead, researchers often reference existing models (e.g. the Capability Maturity Model (CMM)) to shed light on the research concept (e.g. E. S. Andersen & Jessen, 2003; Bibby & Dehe, 2018). These research papers delve into operational aspects of maturity models but do not define maturity nor discuss the components of these models. Before proceeding, it is thus essential to formulate a precise definition that encompasses:

- The meaning of 'maturity' and a 'maturity model',
- The structure and components of a maturity model and
- The potential benefits and applications that maturity models can offer.

Drawing from the work of K. V. Andersen and Henriksen (2006, p.239), the terms maturity and immaturity describe the state of a given level within a continuous process. In the context of software organisations, the processes of an immature organisation are usually improvised by practitioners and their managers during a project (Paulk et al., 1993, p.19). The reactionary nature of these processes also characterises this, meaning that managers are usually focused on addressing immediate crises. On the other hand, mature organisations demonstrate an organisation-wide ability to manage development and maintenance effectively. Besides, mature organisations have a quantitative and objective

basis for assessing product quality and analysing problems with the product and process (Paulk et al., 1993, p.20).

A MM has a sequence of maturity levels that have evolutionary characteristics (Colli et al., 2019), in which levels are successively developed for each concept that requires several capabilities (Schuh et al., 2017). The initial stage is characterised by an organisation having little capabilities in the domain/object under consideration (Hein-Pensel et al., 2023). On the contrary, the final stage represents total maturity. Since no organisation reaches full maturity in the real world, it is logical to discuss a degree of maturity (E. S. Andersen & Jessen, 2003).

Paulk et al. (1993, p.20) define a maturity model as "a specific process to explicitly define, manage, measure and control the evolutionary growth of an entity". Becker et al. (2009) build upon this definition, which we will adopt for our purposes:

> *"A conceptual model consisting of a sequence of discrete maturity levels for a class of processes and represents a desired evolutionary path for these processes"* (Tarhan et al., 2016, p.122)

Important to note is that the literature points to a nuance between maturity models and readiness assessments (Cognet et al., 2023). As mentioned, a maturity model supports an entity to reach a higher level of maturity by following a continuous improvement process. On the other hand, a readiness assessment examines a company's ability to engage in an organisational transformation. These assessments clarify whether an organisation is ready to start the development process (Akdil et al., 2018). Readiness assessments take place before engaging in the maturing process (Schumacher et al., 2016).

Throughout the years, maturity models have been developed as classification schemes within various academic disciplines. In Business Economics, the concept of 'maturity' was applied by Cox (1967) through the Product Life Cycle (PLC). Maturity models were later adopted in IS by Nolan (1979) with his "stages of growth model". However, research experienced a surge in maturity model development following the introduction of the CMM launched by the Software Engineering Institute (Paulk et al., 1993). Since then, numerous maturity models have been created. Some notable examples are the maturity model for Knowledge Management (Kulkarni & Freeze, 2004) and a maturity model for Business Process Management (de Bruin & Rosemann, 2005).

Maturity Models are often structured as a gradual process consisting of multiple stages ordered sequentially (Hein-Pensel et al., 2023). The components present in maturity models are the domain, maturity levels, and dimensions (Bley et al., 2020). The domain is the field of interest on which the maturity model is developed, while the dimensions are the subdivision of an organisational structure into areas of interest. Dimensions are often referred to as capabilities as well.

Maturity Models are primarily used to assess the current state of an organisation, identify areas for improvement, and monitor the progress of implementing these improvements (Iversen et al., 1999; Reis et al., 2017). Pöppelbuß and Röglinger (2011, p.3) further categorise maturity models into three distinct types:

- **Descriptive**: models determining the object's actual state (as-is).
- **Prescriptive**: normative models (to-be) that provide clear recommendations for actions and guidelines for development.
- **Comparative**: models that enable companies and organisations to be located and compared internally and externally.

Despite the benefits of an MM, the model also faces criticism for oversimplifying reality, lacking empirical foundation, and its tendency the neglect the existence of multiple equivalent advantageous paths (de Bruin & Rosemann, 2005; Teo & King, 1997). Moreover, bridging the "knowing-doing gap" is sometimes difficult when an MM does not describe how to effectively perform the improvement actions (Pfeffer & Sutton, 1999).

2.2 Artificial Intelligence

To understand the societal impact of Artificial Intelligence (AI) and evaluate how to responsibly design, implement, and utilise AI, it is crucial to comprehend its components. This section explores the core elements of AI, with a particular focus on GenAI as an evolving domain in the subsequent section. As public organisations explore using Large Language Models (LLMs) to interact with citizens, this topic becomes especially relevant to this thesis.

Generative AI is expected to inject an estimated 2.6 to 4.4 trillion dollars into the global economy annually, making it a highly relevant research field (Chui et al., 2023). The requirements and several prevalent generative models are discussed to gain a comprehensive understanding of this technology.

2.2.1 Defining Artificial Intelligence

AI is a term that represents a broad spectrum of methods and applications. It is not a singular concept, but a collective term for various methods. Figure 2.1 illustrates the different subsets. There are six main categories, each with its specialisations.

Defining AI is challenging due to the field's breadth and the diverse definitions that have emerged. It is a large scientific field with roots in computer science, philosophy, mathematics, psychology, cognitive science and many other disciplines (Dignum, 2019).

Haenlein and Kaplan (2019, p. 5) define AI as "a system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation". Another classic explanation is that every aspect of learning or intelligence feature can be described so that a machine can simulate it (McCarthy et al., 2006). Dignum (2019, p. 10) considers AI the discipline that studies and develops computational artefacts that exhibit some facets of intelligent behaviour.

This research has other objectives than pinning down a precise definition of AI. As the field matures, more unified definitions will follow. Given this research's focus on the

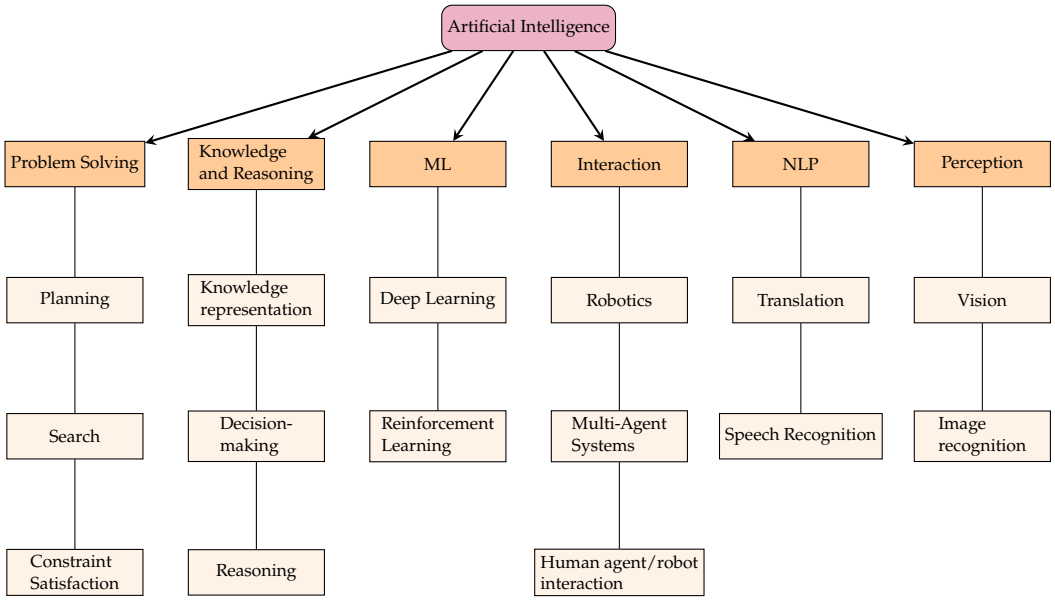


Figure 2.1: Main streams of AI (Dignum, 2019)

public sector, the definition provided by the European Parliament (2024) in the newly adopted AI act is used. They define an AI system as any AI component, be it software or hardware that is a:

> *"machine-based system 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"* (Article 3).

Especially the aspect of the definition that emphasises prediction and decision outputs is crucial when considering Responsible AI.

2.2.2 Generative Artificial Intelligence

Generative AI is a type of Artificial Intelligence technology specialising in producing various types of content such as text, images, audio, and synthetic data. It is a subset of Machine Learning (ML) in which models are trained on large datasets humans generate to recognise underlying statistical patterns. Unlike traditional AI, which is primarily rule-based or deterministic, GenAI operates on probabilistic models to produce novel outputs that are not confined to pre-established patterns.

Organisations anticipate the adoption of GenAI to surge, propelled by the development of new user interfaces that drive its popularity and accessibility. Common use cases of GenAI include:

- **Image generation**: Generative models, particularly General Adverserial Networks (GANs), have showcased remarkable image-generation capabilities. These models

not only revolutionise how we create images but also enhance the quality and functionality of images. Dall-E and Midjourney stand out as prominent examples.

- **Text generation**: Autoregressive models, such as the transformer architecture, can summarise, label, translate and write text. On top of that, they are also able to process audio and speech. The most well-known examples of this use case are OpenAI GPT-4 and Meta LLama.

- **Data augmentation**: Another valuable contribution of generative models is their ability to generate new training examples and augment the existing data. Variational Autoencoders (VAEs) and GANs are most commonly used in this case.

- **Simulation and forecasting**: Generative models can simulate complex systems and predict future behaviour, making them highly valuable in forecasting. For example, GANs have been used for predicting financial time series, as noted by Vuletić et al. (2024).

GenAI leverages various models to enable the creation of new and original content. Among the most prevalent GenAI models are GANs, VAEs, and Transformer models. Their architectural components and training methods are summarised in Table 2.1.

Model	Architecture Components	Training Method
Generative Adversarial Networks	Generator-Discriminator	Adversarial
Variational Autoencoders	Encoder-Decoder	Variational Inference
Transformers	Encoder-Decoder	Supervised

Table 2.1: Architecture components of generative models (Bandi et al., 2023)

A **Generative Adversarial Network** is a framework for estimating generative models through an adversarial process. This process involves the concurrent training of two neural networks: a generative model G that captures the data distribution and a discriminative model D that estimates the probability that a sample originated from the training data as opposed to being produced by G (Goodfellow et al., 2014).

The **discriminator** evaluates real and generated images from the training data to determine their authenticity. Its output, $D(x)$, represents the probability that the input x is real. It aims for $D(x) = 1$ for real images and $D(x) = 0$ for generated ones. Through this process, the discriminator identifies features that contribute to real images. Conversely, the **generator** creates new data instances (such as images) from random noise. We want the generator to create images with $D(x) = 1$ and thus match the real image. Figure 2.2 gives an overview of this process.

Both networks are trained in alternating steps, engaging them in a competition that drives their self-improvement. As the training progresses, the discriminator becomes increasingly adept in spotting tiny differences between the real and generated images and

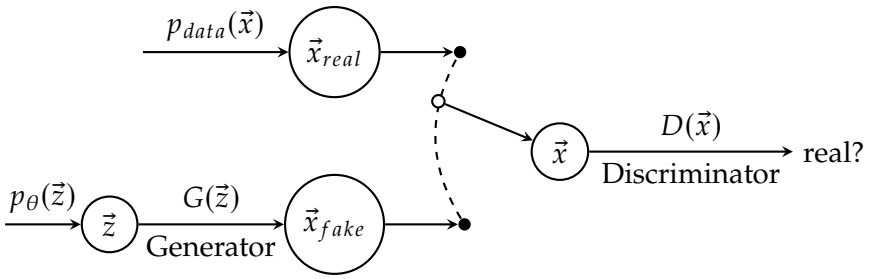


Figure 2.2: Basic scheme of a Generative Adversarial Network

the generator creates images so convincing that the discriminator is unable to distinguish them from the real ones. This dynamic is often conceptualised as a minimax game in which generator G wants to minimise the value function V, while the discriminator D seeks to maximise it:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)]] + \mathbb{E}_{z \sim p_z(z)}[1 - \log D(G(z))]$$

In this equation, $p_{data}(x)$ represents the real data distribution, and $p_z(z)$ corresponds to the noise distribution used by the generator. This equation encapsulates the essence of the adversarial training process, in which the generator and discriminator continuously learn from each other to improve their performance.

Variational Autoencoders (VAE), introduced by Kingma and Welling (2013), represent a class of autoencoders that integrate variational inference within an encoder-decoder structure.

Autoencoders are neural network models primarily used for dimensionality reduction. The encoder, the initial half of the process, transforms raw input data into a compact latent representation. Conversely, the objective of the decoder is to reconstruct the input data as accurately as possible. However, a fundamental limitation of vanilla autoencoders for generation is that the latent space the inputs are converted to, may not be continuous.

Instead of mapping an input to a fixed point in the latent space, VAEs are, by design, continuous, allowing easy random sampling and interpolation. Rather than outputting and encoding vector of size n, it outputs two vectors of size n: a vector of means μ and another vector of standard deviations, σ. This process is also visualised in Figure 2.3. The decoder is defined by $p_0(x|z)$, describing the distribution of the decoded variable given the encoded variable.

The VAE covers a certain area centred around the mean value with a size corresponding to the standard deviation. A sample from anywhere in the area will be similar to the original input. The encodings are thus clustered together more. If a space has more discontinuities (i.e. gaps between clusters) and you want to generate a variation from the input, the decoder will generate an unrealistic output.

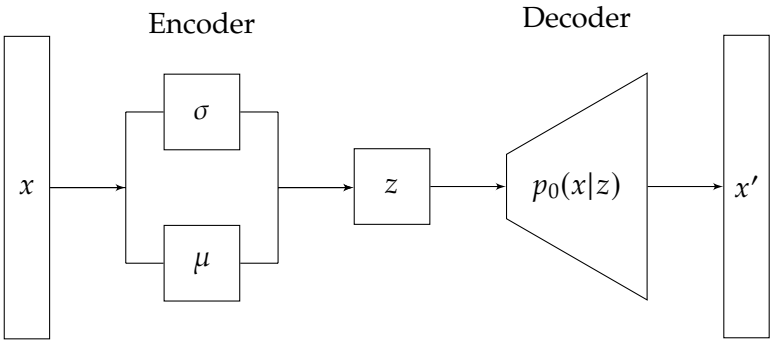


Figure 2.3: Basic scheme of a Variational Autoencoder

The **transformer architecture**, akin to the variational autoencoder, employs an encoder-decoder architecture. Its architecture supports a significant improvement in the performance of deep learning Natural Language Processing (NLP) translation models. The transformer architecture is a fundamental building block of all Large Language Models (LLMs).

Introduced in the seminal paper "Attention is All You Need" by Vaswani et al. (2017, p. 2), the transformer model is distinguished as the first transduction model that exclusively depends on self-attention to compute a representation of its input and output, without using sequence-aligned Recurrent Neural Networks (RNNs) or convolution. The main drawback of RNNs or convolutional layers is their difficulty handling dependencies in long sentences, where related words may be spread far apart. In a sentence such as "the generative model did not know the answer because it lacked training", it is difficult for an algorithm to understand that "it" is associated with the "generative model".

At the heart of the transformer model lies the **self-attention** mechanism, which enables the model to focus on different parts of the input sequence concurrently when making predictions. It takes three vectors as arguments: a query Q, a key K and a value V. It first calculates the dot product of the query and the key, followed by a normalisation of the function through division by $\sqrt{d_k}$. The resulting softmax score, which lies between 0 and 1, is then multiplied by the value factor to obtain an attention score. Following this mechanism, the model can have a better understanding of language.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

2.2.3 Ethical implications Artificial Intelligence

Based on the architectures and components of these generative models, there are several ethical implications useful to take along in this research:

- **Bias and discrimination**: Generative models mirror the data fed. Consequently, if trained on biased datasets, they will inadvertently perpetuate those biases. Especially

since generative models such as GANs have a clear feedback loop. Therefore, it is important to focus on the diversity of the dataset used for training.

- **Lack of transparency**: The lack of transparency in LLMs poses significant challenges to developers and users. Given their complexity, with sometimes billions of parameters, understanding the underlying logic of their decision-making process is not straightforward.
- **Privacy and security concerns**: Significant privacy concerns exist with generative models. These models may inadvertently generate outputs containing sensitive or confidential information gleaned from their training data.
- **Accountability**: The nature of generative models is that they can produce new outcomes. However, it is unclear who is responsible if a generative model produces harmful or misleading content.

The complexity of generative models highlights the importance of adopting a multidisciplinary approach in developing the Responsible AI maturity model. Ethical considerations must be integrated at every stage of the AI lifecycle, from data collection to model deployment. Moreover, it is important to not only examine the data itself but also engage with citizens to continuously re-evaluate the model. Despite the model appearing to provide correct results, the inherent lack of transparency and the clear feedback loop in generative models demand corresponding measures.

2.3 Responsible AI

2.3.1 Defining Responsible AI

Artificial Intelligence holds the potential to improve public services significantly, but it is also coupled with a range of ethical concerns that must be navigated. Therefore, it is integral to consider the concept of ethics alongside the technical aspect of implementation.

The necessity for ethical considerations in new technologies was identified as early as 1985. Moor (1985) highlighted a "policy vacuum" - the absence of established guidelines or principles to guide the ethical use of technology.

Within the field of Responsible AI, three main research streams can be identified, each linked to different academic traditions and research fields (Vassilakopoulou et al., 2022). The first stream, primarily rooted in Computer Science, aims to improve the trustworthiness of AI systems by focusing on aspects such as robustness, algorithmic fairness, explainability, and transparency, often referring to *Trustworthy AI* (Li et al., 2023).

The second stream, drawing from human-computer interaction and human-centred design is known as Human-Centred Artificial Intelligence (HCAI). HCAI focuses on amplifying, augmenting, and enhancing human performance to make systems reliable, safe, and trustworthy (Shneiderman, 2020). HCAI systems emerge when designers,

software engineers, and managers adopt user-centred participatory design methods, engaging with diverse stakeholders.

The third paradigm, rooted in philosophy, builds on the foundations of computer and information ethics and is concerned with the adherence to fundamental human values. Scholars in this community often refer to ethical AI or AI ethics when relating ethical values and principles such as transparency and non-maleficence to AI (Jobin et al., 2019).

The definition of Responsible AI will be drawn from Information Systems, the field most relevant to this research. Vassilakopoulou et al. (2022) define Responsible AI as follows:

> *"Responsible AI is the practice of developing, using and governing AI in a human-centred way to ensure that AI is worthy of being trusted and adheres to fundamental human values."*

It is important to note that Responsible AI is not about giving AI responsibilities. It is not a category of AI artefacts that have special properties or can undertake responsibilities. On the contrary, humans are responsible for AI, and Responsible AI is meant for people and organisations to take more responsibility.

2.3.2 Underlying ethical principles

Although this research is rooted in IS research, it is important to understand the underlying ethical principles to make well-considered decisions about what aspects to include in the maturity model.

Following Floridi et al. (2018), five ethical principles form the foundation of AI ethics. Beneficence, non-maleficence, autonomy, and justice are well-known core principles in bioethics. These principles adapt surprisingly well to the ethical challenges posed by AI. Based on a comparative analysis, only explicability, which incorporates both intelligibility and accountability, has to be added as a principle (Floridi et al., 2018). The research of Jobin et al. (2019) echoes the majority of these principles. In a mapping study that analysed the current corpus of principles and guidelines on ethical AI, they found transparency, justice and fairness, non-maleficence, responsibility, and privacy as core ethical principles mentioned in more than half of all sources. Beneficence (41/84) and autonomy (34/84) were also frequently mentioned in the research corpus but excluded as core principles by Jobin et al. (2019).

However, these principles have been regularly criticised for their generality and lack of practical application, raising questions about how these abstract principles can be translated into concrete frameworks and solutions (Qiang et al., 2023). After all, what does it mean to implement transparency in an AI system (Hagendorff, 2020)?

It is necessary to take a step back and look at the philosophical tradition inherent to these principles to understand the roots of these ethical principles and how they ought to be interpreted. To understand where these ethical principles come from, and how they ought to be interpreted, it is necessary to take a step back and look at the philosophical

tradition that is inherent to these principles. Normative ethics is characterised by three major approaches: deontology, virtue ethics, and utilitarianism. As adopted by Jobin et al. (2019) and Floridi et al. (2018), the deontological perspective advocates for a set of rules, duties, or imperatives. Ethical guidelines regarding data and algorithms, postulate a fixed set of universal principles and maxims which technology developers should adhere to, regardless the consequences. Hagendorff (2020) suggests augmenting deontological AI ethics with virtue ethics, focusing on individual characters and attitudes rather than strict adherence to predefined rules. Ethics is no longer a deontologically inspired tick-box exercise but a project of changing attitudes and strengthening personalities. The advantage of this approach is that ethical guidelines will not be perceived as something whose purpose is to stop or prohibit activity but rather do the opposite. It can broaden the scope of action, promote autonomy and freedom, and foster self-responsibility.

2.4 The foundation for a Responsible AI maturity model

This chapter lays the theoretical foundation for developing a Responsible AI maturity model, with a particular focus on generative AI, which is gaining importance in the public sector. The evaluation of generative models shows that their underlying ethical principles, such as transparency and privacy, are well-aligned with those found in philosophical literature. The theoretical background is also intentionally broad, and the discussion of AI in the public sector is excluded from this chapter. The scattered nature of the literature on AI in the public sector makes it beyond the scope of this thesis to include it. Public sector elements will be incorporated in the empirical part of this research. This section primarily aimed to clarify the relevant concepts and definitions that will guide this research, demarcating the research.

3

Research Design

As stated in chapter 1, the procedural model of Becker et al. (2009) is adopted to conduct this research. Figure 1.1, as shown in Chapter 1, illustrates the process followed in this thesis. The first section of this chapter explains the design choice for the methodology of Becker et al. (2009) and how it compares to other procedural models and methodologies. Subsequent sections detail the remaining steps in the design process of the maturity model.

3.1 Procedural Models

The development of maturity models is primarily dominated by conceptual and design-oriented research designs (Wendler, 2012). In a conceptual research design, the development of the maturity model is outlined, but no empirical validation is conducted. On the other hand, Design Science Research (DSR) is a problem-solving paradigm that aims to extend technology and science knowledge bases by creating innovative artefacts that solve problems and improve their environment (vom Brocke et al., 2020). In other words, the artefacts are designed to interact with the problem context to improve it (Wieringa, 2014). This notion implies that an artefact has to be tested, for example, by proof of concept or a case study, to ensure its applicability and benefits. For this research, the design science paradigm is most fitting. The maturity model will be iteratively tested and validated to ensure its practical relevance.

One of the frameworks most frequently cited for conducting design-oriented research is the DSRM by Hevner et al. (2004). They aim to inform IS researchers and practitioners on conducting, evaluating, and presenting design science research. They achieve this by describing the boundaries of IS design science research via a conceptual framework and guidelines for conducting and evaluating good design science research.

These guidelines, as defined by Hevner et al. (2004), serve as practice rules for conducting Design Science research and help to understand the definition and meaning of DSR. However, to fulfil the requirements of a common DSRM, it is necessary to follow specific procedures, including a process model and a mental model. Peffers et al. (2007,

Guideline	Description
Problem Definition	The targeted domain and the target group need to be determined, and the problem relevance must be clearly demonstrated.
Comparison of existing models	Analysis of existing, but unsatisfactory maturity models.
Determination of development strategy	Detailed documentation of the design process of the maturity model. There are four design choices: a completely new model design, the enhancement of an existing model, the combination of several models into a new one, or the transfer of structures or contents from existing models to new application domains.
Iterative maturity model development	Development of a first model with literature research or explorative research methods. Iterations of selecting the design level, selecting the approach, designing the model selection, and test the results.
Conception of transfer and evaluation	The different forms of result transfer for the academic and user communities need to be determined. This can be the publication of document-based check lists and manuals, or a software-tool supported accessibility of the maturity model.
Implementation of transfer media	Make the maturity model accessible for all previous defined user groups.
Evaluation	Establish whether the maturity model provides the projected benefits and an improved solution for the defined problem.

Table 3.1: Procedural model for the development of maturity models (Becker et al., 2009)

p.49) define a methodology as "a system of principles, practices, and procedures applied to a specific branch of knowledge". IS researchers have not previously focused on developing a process and mental model for design science research.

Therefore, Peffers et al. (2007) propose a six-step procedure consisting of problem identification and motivation, definition of the objectives for a solution, design and development, demonstration, evaluation, and communication. It serves as a roadmap for design research.

Since then, several researchers have developed procedural models for design science research. Wieringa (2014) developed a design cycle similar to the DSRM from Peffers et al. (2007). The design cycle involves the problem investigation, treatment design, and treatment validation. In the DSRM of Peffers et al. (2007), the treatment validation is split into development, demonstration and evaluation. Communication is viewed as part of research management in the cycle of Wieringa (2014).

Specifically tailored to the development of maturity models, de Bruin and Rosemann (2005) outlined the primary phases of general model development. The phases included in this model are scope, design, populate, test, deploy, and maintain. The first phase involves determining the model's desired scope, with a crucial decision being the model's focus, which could be either domain-specific or general. When opting for a domain-specific model, an extensive review of existing literature in the respective domain, related

Common	Peffers et al. (2007)	Wieringa (2014)	de Bruin and Rosemann (2005)	Becker et al. (2009)
Scope	Problem identification	Problem investigation	Scope	Problem definition
	Objectives of a solution	Treatment design		
Design				Comparison of existing maturity models
	Design...	Treatment design: the rest	Design	Determination of development process
Develop	...and development	Validation: instrument development	Populate	Iterative maturity model development
Implement	Demonstration	Validation: effects, tradeoffs, sensitivity?	Test	Conception of transfer and evaluation
	Evaluation	Validation: do effects satisfy requirements?	Deploy	Implementation of transfer media
	Communication		Maintain	Evaluation

Table 3.2: Mapping of procedural models

domains, and maturity models could be conducted to gain a profound understanding of historical and contemporary domain issues. In the design phases, the intended audience's needs should be incorporated, reflecting why they seek to apply the model, how it can be applied to varying organisational structures, who needs to be involved in its application, and what outcomes can be achieved through its implementation.

During the populate phase, the model's content is determined, requiring a clear understanding of what aspects need to be measured in the maturity assessment and how these aspects can be measured. Once the model is populated, rigorous testing is essential to ensure the model's relevance and reliability. Subsequently, the deployment phase involves making the model available for use and validating its generalisability. Finally, the model must be maintained, allowing it to evolve as domain knowledge deepens and understanding broadens.

This procedural model from de Bruin and Rosemann (2005) has been further refined by Becker et al. (2009), as shown in Table 3.1. Table K.5 shows the mapping results of the different methodologies. Each relevant methodology is mapped to the common process phase as defined by van Steenbergen et al. (2010). For this research, the procedural model of Becker et al. (2009) will be followed in combination with the design cycle of Wieringa (2014). The former methodology will lead, but Wieringa (2014) has a more extensive description of developing and testing an artefact.

3.2 Research Model

The research process, as illustrated in figure 3.1 follows the common phase distinction made by van Steenbergen et al. (2010). Chapters 1 and 2 have already addressed the problem definition and related work, while the remaining chapters will focus on the design, development, and implementation of the maturity model.

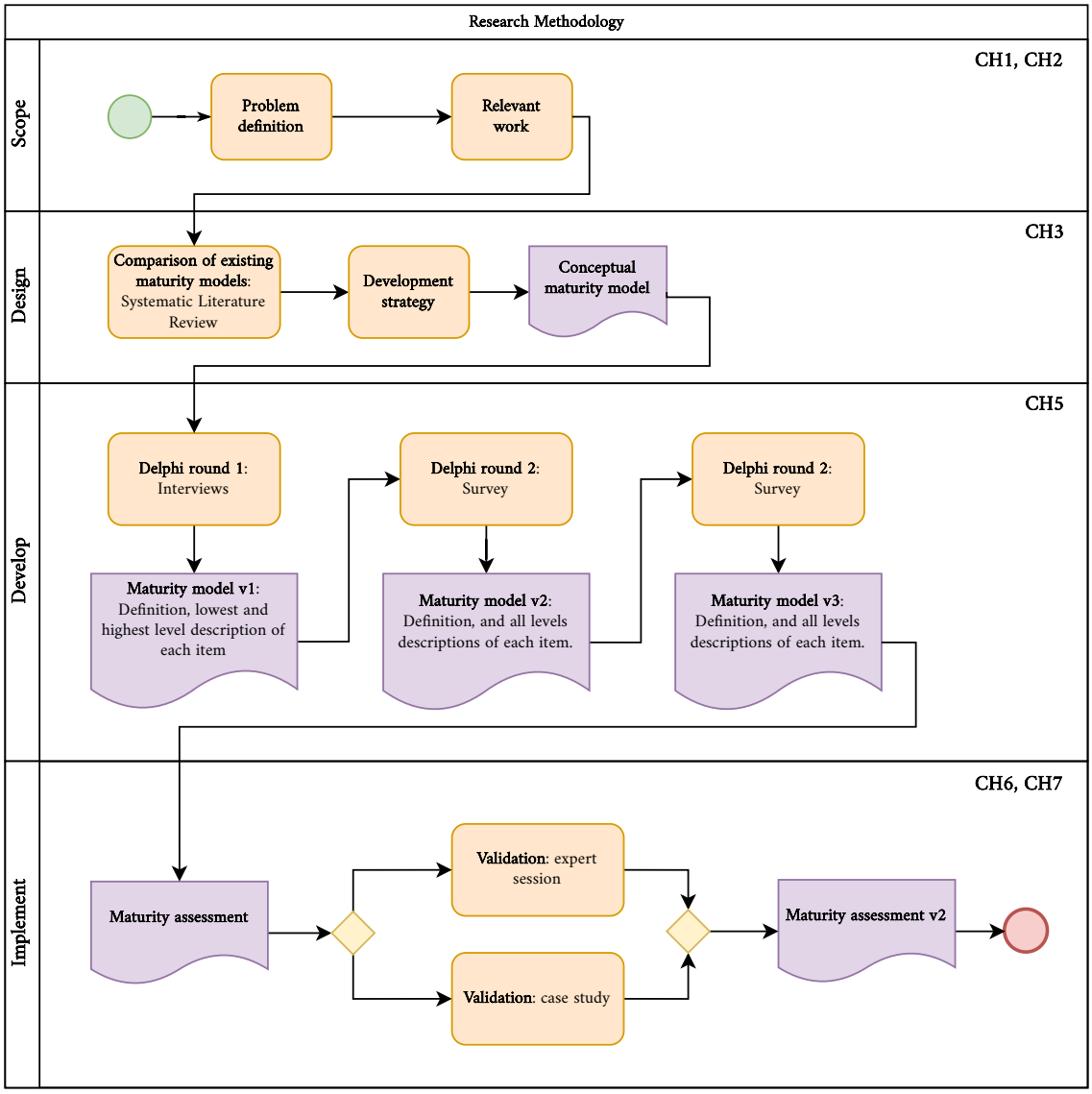


Figure 3.1: Research Model

3.3 Literature Review

To justify the development of a new maturity model, it is essential to conduct a thorough review of the current models available in academic literature, following step two of Becker et al.'s 2009 procedural model, which involves the **comparison of existing models**. A Systematic Literature Review (SLR) is performed to find all relevant models. Based on

the findings of the literature review, it will become evident whether there is a need to develop an entirely new maturity model, enhance an existing one, or integrate various models into a novel model (i.e. **determination of development strategy**). The literature review will result in a conceptual model, which will then be tested and validated through a Delphi study, as outlined in the next section. The methods of Wolfswinkel et al. (2013) and Page et al. (2021) are used to identify the relevant literature. The detailed procedure is described in Chapter 4.

3.4 Delphi Study

For the **iterative development** of the maturity model (step 4 of procedural model), the Delphi method has been selected. The Delphi method is a structured, organised, and iterative process designed to distill and correlate opinions from an expert panel concerning a particular problem, topic, or task (Alarabiat & Ramos, 2019). Originating from a series of studies by the RAND Corporation in the 1950s (Okoli & Pawlowski, 2004), the Delphi method aims to obtain the most reliable consensus of expert opinions through a series of intensive questionnaires interspersed with controlled opinion feedback (Rowe & Wright, 1999).

According to Rowe and Wright (1999, p.354), four key features define a Delphi study:

1. **Anonymity**: this characteristic promotes panelist independence and allows group participants to express their judgements individually and avoid undue social pressures (Rowe & Wright, 1999; Skinner et al., 2015). It also eliminates the potential for participants to mimick others (Skinner et al., 2015).

2. **Iteration**: a Delphi study consists of several iterations of the questionnaire. Individuals are given the opportunity to change their opinions and judgements after each iteration without fear of losing face in the eyes of the others in the group (Rowe & Wright, 1999).

3. **Controlled feedback**: Between each round, participants are informed on the thoughts of their anonymous fellow participants (Strasser, 2017). The collected opinions are analysed, and information on the answers is provided to the panelist for comments or to guide the next round (Skinner et al., 2015).

4. **Statistical aggregation of group response**: the group opinion is defined as an appropriate aggregate of individual opinions on the final round (Dalkey, 1969). This ensures that the final results reflect the collective judgement of the panel, rather than being influenced by any single participant.

3.4.1 Design Choice

According to Skinner et al. (2015), the Delphi method is particularly appropriate for acquiring expert recommendations when addressing an IS research issue. Skulmoski et al. (2007) also mentions that the Delphi method is a suitable candidate for research projects in the IS domain when there is incomplete knowledge about phenomena.

The selection of a Delphi study for this research is motivated by its effectiveness in developing and validating new maturity models (de Bruin & Rosemann, 2007; Martinek-Jaguszewska & Rogowski, 2022). In Information Systems research, the methodology has been primarily used for forecasting and issue identification, and framework development (Okoli & Pawlowski, 2004), fitting the purpose of this study. It is proven to be a suitable method for research that is exploratory (de Bruin & Rosemann, 2005). Responsible AI is still an evolving research field and researchers are still struggling to make the translation from the 'what' to the 'how' of Reponsible AI.

Compared to a focus group, a Delphi study has several advantages. The anonymity of the method leads to more creative outcomes (Okoli & Pawlowski, 2004), and issues inherent in face-to-face meetings groups such as dominant personalities, conflict and group pressures are virtually eliminated (de Bruin et al., 2005). Especially, in the Netherlands, a low-context culture, where opinions and even negative feedback tend to be stated openly, anonymity could lead to less biased results.

Similarly, the Delphi study also provides some advantages over traditional expert interviews. Interviews can generate volumes of messy data, which are difficult to analyse(Brown, 2018). The Delphi study offers a more organised way to process the data and develop the maturity model.

Obviously, several weaknesses and critics of Delphi studies are present. Most frequently mentioned in literature are:

- **Apparent consensus**: critics have argued that consensus is often only 'apparent', and that the convergence of responses in mainly attributable to other social-psychological factors leading to conformity (Rowe & Wright, 1999). It is perceived to force consensus and is weakened by not allowing participants to discuss the issues raised (Hasson et al., 2000).

- **Response rate**: a Delphi study is time-consuming and laborious for both researchers and participants, and thus vulnerable to drop-outs. There is long temporal commitment, distraction between rounds, or disappointment with the process (Donohoe & Needham, 2008).

- **Generalisability or external validity**: The Delphi method employs a non-representative sample of experts to form opinions on complex, multi-disciplinary problems. Generalising the opinions and estimations of such a non-representative group to a larger population can be problematic at best (Worrell et al., 2013).

- **Expert opinion**: In relation to the previous point, Delphi studies rely purely on expert opinion to generate findings. There is a potential for bias in the selection as the exact composition of the panel can affect the results obtained (Keeney et al., 2001). Simply because individuals have knowledge of a particular topic does not necessarily mean that they are experts.

To address these limitations and ensure validity and reliability of the results, several measures are taken. Firstly, to reduce the risk of illusory expertise and to systematise the process for identifying experts, a vetted Delphi sampling technique, known as the Knowledge Resource Nomination Worksheet (KRNW) sampling procedure is used (Parrish & Sadera, 2018). This approach helps to overcome expert opinion and generalisability limitations. By assembling a diverse group of experts, the validity for generalising the results to a broader population is increased. Additionally, a clear definition of when consensus is reached, how to measure it, and the communication about it towards participants will contribute to overcome a fake consensus and high dropout rates.

3.4.2 Delphi panel selection

To ensure the validity and relevance of the Delphi results, the right experts need to be selected. The purpose of the KRNW is to categorise experts before identifying them individually, to avoid overlooking any important class of experts (Okoli & Pawlowski, 2004). The first step is to prepare the KRNW by identifying relevant disciplines, skills, and organisations. The following categories were identified as relevant:

- **Academics**: Scholars in the field of Human-Computer Interaction, Computer Science, Information Systems, or Philosophy, who could provide well-founded insights on Responsible AI. These experts will be identified through their publications in the field.

- **Consultants**: Experts with hands-on experience in Responsible AI projects within the public sector. These experts are selected based on relevant project experience.

- **Employees**: Professionals from public organisations with practical experience in implementing Responsible AI practices, such as data scientists or ethical officers.

3.4.3 Delphi panel size

The validity, efficacy, and reliability of Delphi study results are influenced by the size of the expert group, as noted by Donohoe and Needham (2008). There is an ongoing debate regarding theoptimal panel size for Delphi studies. Okoli and Pawlowski (2004) stresses that Delphi sample sizes depend more on group dynamics in reaching consensus than on statistical power. There is thus no standard for what constitutes a small or large panel. Following other studies in the IS research community, the majority of studies report an

initial panel size of 14 to 30 (Paré et al., 2013). Alarabiat and Ramos (2019) note that when experts have consistnet and extensive experience, a panel of 10 to 15 member is sufficient.

3.4.4 Delphi rounds

The Delphi process consists of three rounds to gather and refine expert opinions. Researchers highlight the exploratory nature of the first round, with some choosing a qualitative approach to allow experts to generate ideas and express their views (Okoli & Pawlowski, 2004).

The goal of **Round 1** is to investigate the initial model's structure, main dimensions, and items. To gain deeper insights into the experts' opinions, the first round consists of interviews, allowing participants to more openly explain their reasoning behind the most important dimensions and items.

The survey in **Round 2** includes the updated dimensions and items and asks the opinion of the experts on the lowest and the highest maturity levels for each of the items. Before sending out the survey, pilot testing with one individual preceded the implementation, as suggested by the literature (Hasson et al., 2000).

Round 3 provides the experts with the complete model. It includes the dimensions, items and the five maturity level for each of the items. It provides a final opportunity for participants to revise their judgements.

3.5 Maturity Assessment

To facilitate the **conception of transfer and evaluation**, and the **implementation of transfer media**, an assessment tool will be created in Excel. For every dimension, a participant can fill out questions to determine their maturity score. This will generate a spider chart that illustrates the organisation's maturity progress, benchmarked against other organisations in a similar sector.

There is limited research on the development of maturity assessment tools. Fukas et al. (2023) developed an AI maturity assessment tool, highlighting the need for a survey-based assessment. Therefore, the assessment tool will also be survey-based. The following functional and non-functional requirements have been formulated to guide the design process of the Excel sheet:

- **Functional requirements**: (1) The tool should automatically calculate maturity scores based on user responses, and (2) the tool should generate a spider chart displaying the current and desired maturity levels.

- **Non-functional requirements**: (1) The maturity assessment should be completable within the duration of an interview, and (2) the tool should be designed to allow users to intuitively perform an assessment.

The design and development of the assessment tool is not the primary goal of this research. Therefore, only a limited set of requirements has been included.

3.6 Evaluation

The **evaluation** of the designed maturity model consists of two components:

1. Assessing the value of the assessment through an expert session, and
2. applying the maturity assessment to a sample of relevant public organisations.

These components align with a Type II and Type III evaluation, as defined by Helgesson et al. (2011). A type II evaluation involves practitioners who have not been involved in developing the maturity model, and a type III evaluation is conducted in which the maturity model is used in a practical setting.

3.6.1 Evaluation criteria

First, there needs to be an understanding of the "what" of evaluation (Prat et al., 2015). It means that the evaluation criteria must be defined to conduct an appropriate evaluation. In IS research, models such as Technology Acceptance Model (TAM) and Unified Theory of Acceptance and Use of Technology (UTAUT) are often used to understand and predict the acceptance of a design artifact (Venkatesh et al., 2003). However, these models have faced criticism for their narrow focus and potential lack of applicability across different contexts.

While these models are relevant for examining the maturity model, most studies applying TAM or UTAUT typically involve surveys of hundreds of users. This research does not aim to replicate such a survey-based approach to that extent. Additionally, the evaluation criteria are not exhaustive, as there are more factors to consider beyond ease of use and understandability to evaluate the maturity model.

Therefore, a deliberate choice is made to gather evaluation criteria from different sources, selecting the most relevant ones for evaluating the maturity assessment. The taxonomy as described by Prat et al. (2015) is used to find evaluation criteria that should be included. Most of the questions were adopted from Salah et al. (2014). Their article already provided evaluation criteria specifically tailored to maturity models. The evaluation criteria that were considered important were ease of use, usefulness, understandability, and completeness. *Ease of use* and *usefulness* were included to predict the intention to use the artifact.

Additionally, *understandability* and *completeness* were included as evaluation criteria to indicate whether filling out the model indeed leads to a better understanding of your responsible AI maturity. The participants in expert sessions were asked to fill out a survey with the statements as shown in table 3.3. On top of that, these evaluation criteria were also discussed during the interviews and expert sessions.

Criteria	Definition	Survey question
Ease-of-use	The degree to which the artifact can be comprehended, both at a global level and at the detailed level of the elements and relationships inside the artifact (Prat et al., 2015, p.257).	(1) The documentation is easy to use, (2) The maturity model is easy to use, (3) The assessment is easy to use (Salah et al., 2014).
Understandability	The degree to which an individual believes that using a particular system would be free of physical and mental effort (Davis, 1989, p.26).	(1) The documentation is understandable, (2) The maturity model is understandable, (3) The assessment is understandable, (4) The dimensions and items are understandable (Salah et al., 2014).
Usefulness	The degree to which the artifact positively impacts the task performance of individuals (Prat et al., 2015, p.266)	(1) The maturity model is useful for conducting assessments, and (2) The maturity model is practical for use in industry
Completeness	The degree to which the structure of the artifact contains all necessary elements and relationships between elements. (Prat et al., 2015, p.266)	The maturity model assessment criteria cover all the relevant aspects of responsible AI

Table 3.3: Evaluation criteria maturity model

3.6.2 Type II: Expert evaluation

A group of academics and consultants are included as the expert group. For both groups, the research is presented at the start of the session to provide further detail about the development process and the study's outcomes. The presentation aims to enhance interaction during the session, enabling them to understand the model more thoroughly. Following the research presentation, the experts will be asked about the model's usefulness, ease of use, understandability, and completeness. These evaluation criteria, which are further elaborated in chapter 7, form the basis of the discussion. The remainder of the discussion will be open-ended.

3.6.3 Type III: Case study evaluation

Two organisations will be evaluated through a case study. Participants are asked to fill out the maturity assessment before the interview. They are encouraged to discuss specific questions with colleagues if they are unsure. In the assessment, they not only fill out the organisation's current maturity but also the desired future state. Filling out the desired state creates an overview of where different public organisations aspire to be and serves as a benchmark. Besides filling out the maturity assessment, participants are also asked to

rate the usefulness, ease of use, understandability, and completeness, similar to the expert evaluation.

Following the assessment, participants are interviewed to share their views on the model and to highlight any missing or redundant aspects. This interview also provides an opportunity to delve deeper into the usefulness and practicality of the maturity assessment.

4

Systematic Literature Review

A general literature review is conducted to develop a comprehensive understanding of existing AI maturity models in academic literature. It would be preferable to conduct a SLR, which has the advantage of more objective and transparent data collection (Tranfield et al., 2003). However, multiple researchers would be required to ensure the validity and reliability of the performed review.

A structured approach is followed for this literature review to enhance transparency and minimise bias. The review structure is based on the grounded theory approach proposed by Wolfswinkel et al. (2013). With the approach, inclusion and exclusion criteria, as well as search terms, are defined before searching and selecting relevant articles. The process concludes with coding the articles. Additionally, the PRISMA flow chart adapted from Page et al. (2021) represents the selection of articles visually and logically for the literature review. The flow chart can be found in Appendix A.

The literature review aims to align with the procedural model of Becker et al. (2009), as discussed in the previous chapter. According to the second step of the model, a comparison with existing maturity models should be made to determine the development strategy of a new Responsible AI maturity model.

	Inclusion Criteria	Exclusion Criteria
Topic		**Ex1:** Publication does not include a maturity model or readiness assessment
	In1: Publication is related to Responsible AI	**Ex2:** Publication is not related to a maturity model for Artificial Intelligence
Language		**Ex3:** Publication is not written in English
Availability		**Ex4:** Full text of the publication is not available
Publication date		**Ex5:** Publication has been published before 2019
Publication outlet	**In2:** Journals, Conference proceedings, and Book chapters	**Ex6:** Grey literature

Table 4.1: Inclusion and Exclusion criteria

4.1 Literature Methodology

4.1.1 Defining the literature review

To ensure that only relevant articles are included in the review, inclusion and exclusion criteria are delineated to guide the selection of articles. The scope of the review is confined to academic articles, including conference proceedings, journal articles, and book chapters. This decision aligns with the findings from recent literature reviews by Sadiq et al. (2021) and Akbarighatar (2022), which indicate a proliferation of AI maturity models in recent academic articles. Therefore, it is unnecessary to include grey literature and practical frameworks in our search. Additionally, the scope is restricted to publications from 2019 onwards, considering the significant changes in AI in recent years. Interestingly, even when looking at older criteria, none of the papers before 2019 met all the criteria to be included in the review. This statistic also confirms the thought of the fast-changing subject field. The criteria also stipulate the exclusion of articles that do not present a (preliminary) maturity model or those that feature a maturity model unrelated to AI. The criteria are defined in Table 4.1 to facilitate a transparent and replicable selection process.

Sources	Search string
Scopus	("Artificial Intelligence" OR "AI") AND ("Maturity Model" OR "Maturity Assessment")
Web of Science	: ("Artificial Intelligence" OR "AI") AND ("Maturity Model" OR "Maturity Assessment")
Google Scholar	allintitle: Maturity "AI" OR "Artificial Intelligence"

Table 4.2: Database search strings

4.1.2 Sources and search terms

Databases included in the search are Scopus, Web of Science and Google Scholar to get a wide overview of the available literature and ensure no relevant articles are excluded. Table 4.2 shows the search terms that have been used for every database. The search string of Google Scholar is more detailed than that of Scopus and Web of Science because the search initially resulted in too many articles being returned. The PRISMA chart in Appendix A shows how the articles have been selected in more detail.

4.1.3 Coding

This research uses a deductive coding scheme, as detailed in Table 7.1. Since no uniform and validated method for comparing maturity models exists, a new classification scheme has been developed. This scheme is based on a combination of papers, including the previous literature review by Sadiq et al. (2021).

As Webster and Watson (2002) suggested, a concept matrix is create to identify the relevant dimensions further. Although creating a concept matrix is systematic, it involves subjective interpretation to determine which dimensions are most important for the new maturity model.

Class	Subclasses	Description
Purpose of use	Descriptive	It is applied for as-is assessments where the current capabilities of the entity under investigation are assessed with respect to given criteria (Pöppelbuß & Röglinger, 2011).
	Prescriptive	It indicates how to identify desirable maturity levels and provides guidelines on improvement measures (Pöppelbuß & Röglinger, 2011).
	Comparative	It allows for internal or external benchmarking. Given sufficient historical data from a large number of assessment participants, the maturity levels of similar business units and organizations can be compared (Pöppelbuß & Röglinger, 2011).
Typology	Structured models	A formal and complex structure, similar to the CMM (Correia et al., 2017; Fraser et al., 2003)
	Maturity grids	A number of maturity levels attending to the several aspects of the research area (Correia et al., 2017)
	Likert-scale questionnaire	A set of questions where the respondent classify the company or SC performance on a scale from 1 to n (Correia et al., 2017)
	Hybrid models	A combination of characteristics of maturity grids and Likert-like model structure (Correia et al., 2017)
Architecture	Staged	A cumulative set of areas defining each level. All the areas included in a level need to be successfully achieved before moving to the next level (Correia et al., 2017).
	Continuous	A set of areas that can be approached separately. Rather than having to address all the areas for a given level, the focus of improvement can be a specific area (Correia et al., 2017).
	Others	Other representations not included in previous subcategories.
Model focus	General	The maturity model has been developed for general application (de Bruin & Rosemann, 2005).
	Domain specific	The maturity model has been developed for a specific domain

Table 4.3: Components to compare maturity models

4.1.4 Previous literature reviews on AI maturity models

Table 4.4 shows a summary of previous literature reviews on AI maturity models. Despite these reviews, there remains a clear gap that the current literature review aims to address. Firstly, it provides an update on the newest AI maturity models published, as neither existing review includes articles from 2023 and 2024. Additionally, this literature review

has a different focus compared to previous reviews. While the primary purpose of the previous reviews is to gain insight into the research field, methodologies, and characteristics of AI maturity models, the current review centres on constructing a new Responsible AI maturity model based on existing models. The field remains dispersed, and one goal of this review is to consolidate existing models and leverage their strongest aspects.

Study	Purpose	Yrs	# studies
Sadiq et al. (2021)	Allow a deeper understanding of the methodological issues relevant to maturity models, especially in terms of the objectives, methods employed to develop and validate the models, and the scope and characteristics of maturity model development	2015-2020	15
Reichl and Rudolf (2023)	Present capability, readiness, and maturity models related to artificial intelligence and provide them with detailed explanations.	2018-2022	16
Akbarighatar (2022)	Provide an overview of extant research on maturity and readiness models for RAI and provide a comprehensive model leveraging this prior research	2017-2022	35

Table 4.4: Previous literature reviews

4.2 Literature Characteristics

This chapter synthesises relevant literature on maturity models for AI ethics. As discussed in Chapter 3, a literature review helps understand existing maturity models and decide whether to design a new model, extend an existing one, or combine different models.

The review examines all AI maturity models developed in academic literature, excluding grey literature. Both general AI maturity models and those specifically aimed at Responsible AI are included to provide a comprehensive overview of foundational AI and Responsible AI capabilities. As Jantunen et al. (2021) indicated, there is a risk that focusing on Responsible AI too much can draw the focus away from technical aspects. This focus on Responsible AI could lead to a situation in which the maturity model would face issues in practical application, similar to existing guidelines.

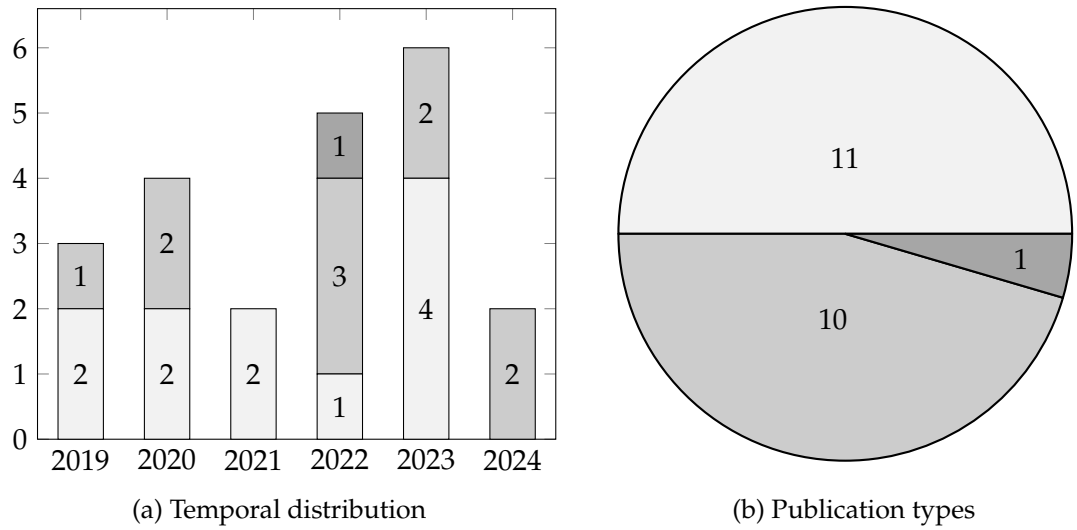
Therefore, a sociotechnical approach is adopted for the maturity model. Most models focus on technical or social aspects, but research shows the need to address both (Akbarighatar, 2022; Asatiani et al., 2021, e.g.). Advocates of a sociotechnical approach argue that it is important to consider both the technical artefacts and the individuals or collectives that develop and use these artefacts within social contexts (Asatiani et al., 2021),

In this chapter, the results of all identified literature are discussed, highlighting key findings and insights. The characteristics of existing models are analysed, and how they can contribute towards the new Responsible AI maturity model is explored. This synthesis

aims to consolidate the field and leverage the strongest aspects of current models to create a conceptual maturity model for Responsible AI.

4.2.1 General characteristics of the studies

In recent years, there has been a growing interest in AI maturity models, as can be seen in figure 4.1a. Notably, 2022 and 2023 witnessed a significant increase in publications related to this topic, highlighting the current relevance of the topic. However, the field remains relatively new and unexplored. Approximately half of the included articles are from conference proceedings, indicating ongoing research and development (see figure 4.1b. Authors of these conference papers, such as Alsheibani et al. (2019) and Hartikainen et al. (2023), have proposed preliminary maturity models with plans for further validation. Considering that the number of conference proceedings also means that some articles lack peer review, it is essential.



(a) Temporal distribution

(b) Publication types

Figure 4.1: Characteristics of previous studies

4.2.2 Maturity model components

In examining existing maturity models, a comparison is made regarding the maturity levels and dimensions they utilise (as shown in Table 4.5). The findings reveal that most models incorporate maturity levels ranging from 3 to 5. Fewer maturity levels offer clear advantages for model construction, allowing for a clearer definition of the developmental path. On the other hand, a larger number of maturity levels offers a better differentiation between organisations. When it comes to dimensions, there is a wider distribution. Across all models, the dimensions span from 2 to 13, a relatively broad range. By carefully

Author	Maturity levels	Descriptor	Dimensions levels	Name	Descriptors
Akkiraju et al. (2020)	5	Initial, Repeatable, Defined, Managed, Optimizing	9	Capabilities	AI Model Goal Setting, Data Pipeline Management, Feature Preparation Pipeline, Train Pipeline Management, Test Pipeline Management, Model Quality, Performance and model management, Model Error Analysis, Model Fairness & Trust, Model Transparency
Alsheibani et al. (2019)	5	Initial, Assessing, Determined, Managed, Optimise	4	Dimensions	AI functions, Data structure, People, Organisational
Cho et al. (2023)	5	Incomplete, Performed, Managed, Established, Predictable, Optimizing	13	Processes	Software requirements analysis, Software architecture design, Data collection, Data cleaning, Data preprocessing, Training process management, Performance evaluation of AI model, Safety evaluation of AI model, Final AI model management, System safety evaluation, System safety preparedness, AI infrastructure, AI model operation management
Coates and Martin (2019)	5	N/A	11	Constructs	*Design and development stages*: business, people, user, data, algorithm, compliance *Post development stages*: business data, testing, client feedback, compliance
Dotan et al. (2024)	5	(1) LLL, (2) MML, MLL, or HLL, (3) HMM, HHL, HML, or MMM, (4) HHM, (5) HHH	11	Pillars/ Dimensions	*NIST pillars*: Map, Measure, Manage, Govern *Responsibility Dimensions*: Accuracy, Fairness, Privacy, Security, Ecology, IP&Copyright, Human oversight
Ellefsen et al. (2019)	4	Novice, Ready, Proficient, Advanced	5		Strategy, Organisation, Data, Technology, Operations
Ferreira et al. (2023)	4	Unaware Exploratory, Proactive, Strategic	7	Requirements	Human agency & oversight, Technical robustness & safety, Privacy & data governance, Transparancy, Diversity, non-discrimination & fairness, Societal & environmental wellbeing, Accountability

Fukas et al. (2021); Fukas et al. (2023)	5	Initial, Assessing, Determined, Managed, Optimized	8	Dimensions	Technologies, Data, People & Competences, Organisation & Processes, Strategy & Management, Budget, Products & Services, Ethics & Regulations
Hartikainen et al. (2023)	3	N/A	6	Criteria	Explainability, Transparency, Fairness, Accountability, Collaboration & human control, Working with AI's uncertainty
Holmström (2022)	5	None, Low, Moderate, High, Excellent	4	Dimensions	Technologies, Activities, Boundaries, Goals
Jantunen et al. (2021)	5	Ad hoc, Optimized (dimensions in between have not been defined)	9	Requirements	Understanding stakeholders, Accountability, Data privacy, Fairness, Human agency, Safety & security, System oversight, Transparency, Wellbeing
Krijger et al. (2022)	5	N/A	6	Dimensions	Awareness & Culture, Policy, Governance, Communication & Strategy, Development processes, Tooling
Lichtenthaler (2020)	5	Initial intent, Independent initiative, Interactive implementation, Interdependent innovation, Integrated intelligence	N/A	N/A	N/A
Mylrea and Robinson (2023)	4	No control, Partially implemented, Largely implemented, Fully implemented	7	Pillars	Explainability, Data privacy, Technical robustness & safety, Transparency, Data use & design, Societal well-being, Accountability
Noymanee et al. (2022)	5	Rookie level, Beginner level, Operational level, Expert level, Master level	4	Aspects	Strategy, Organisation, Information, Technology
Schaschek and Engel (2023)	5	Being aware, Taking first steps, Approaching strategically, Operationalising, Innovating	4	Dimensions	Data, Technology, People & culture, Processes
Schmidt et al. (2022)	5	Exploring, Experimenting, Formalising, Optimising, Transforming	5	Dimensions	Strategy, Data, Technology, People, Governance
Schuster et al. (2021); Schuster and Waidelich (2022)	5	Novice, Explorer, User, Innovator, Pioneer	7	Dimensions	Strategy, Organisation, Culture/Mindset, Technology, Data, Privacy, Ethics
Sonntag et al. (2024)	5	Initial, Experimental, Practicing, Integrated, Transformed	5	Dimensions	Culture & Competencies, Strategy, Data, Organisation, Processes, Technology

Uren and Edwards (2023)	3	Laying the foundations of AI, Adoption of AI, Mature AI	4	Components	People, Process, Technology, Data
Yams et al. (2020)	5	Foundational, Experimenting, Operational, Inquiring, Integrated	6	Dimensions	Strategy, Ecosystems, Mindsets, Organisation, Data, Technology
Zhobe et al. (2021)	4	The starter, The aspiring, The equipped, The leader	2	Categories	Capabilities, Vision

Table 4.5: Components existing maturity models

comparing these dimensions and assessing their alignment with the ethical principles proposed by Jobin et al. (2019), the relevant dimensions can be distilled for a new maturity model for Responsible AI.

4.2.3 Purpose and scope of each Maturity Model

Appendix C provides a complete overview of the purpose and scope of each maturity model, highlighting the most important aspects. These include whether the model includes only foundational elements or also focuses on Responsible AI.

The maturity models identified in the literature each provide unique perspectives and frameworks tailored to specific aspects of AI. Most of the models are broadly applicable to any organisation. However, four models are tailored to specific applications or industries: Ellefsen et al. (2019) focuses on Logistics and Industry 4.0, Fukas et al. (2021) on auditing, Schmidt et al. (2022) on Solar PV-plant SMEs, and Sonntag et al. (2024) on manufacturing. The maturity model of Noymanee et al. (2022) is the only one to focus on the public sector, but it is not clear in what way it is tailored to the public sector.

Another observation is the focus of these models. Nine models emphasise HCAI Hartikainen et al. (2023, e.g.), Responsible AI Ferreira et al. (2023, e.g.), or AI ethics Krijger et al. (2022, e.g.), while the remaining articles primarily address foundational AI. These results indicate that an extensive Responsible AI maturity model for the public sector is not present in the literature.

4.3 Development of a conceptual model

After synthesising the literature, six dimensions and five levels have emerged, resulting in the first version of a Responsible AI Maturity Model, as presented in Table 4.6 The dimensions are **Strategy**, **Culture & Competences**, **Organisation**, **Governance & Processes**, **Data management**, and **Technology**. The corresponding levels are **Initial**, **Experimental**, **Practicing**, **Integrated**, and **Transformed**. A table with the dimensions and their corresponding references is presented on the next page. Further descriptions of each dimension are detailed in the following subsections.

Dimension	Definition	References	#
Strategy	The overarching vision for how organisations will operate in the future using Artificial Intelligence and the plans on how to communicate the efforts on Responsible AI	Fukas et al. (2021), Holmström (2022), Noymanee et al. (2022), Schmidt et al. (2022), Schuster et al. (2021), Sonntag et al. (2024), and Yams et al. (2020)	7
Culture & Competences	The collective mindset, skills, and training in the context of AI ethics within an organisation. It highlights the need for continuous education and awareness about technical and ethical aspects of AI	Alsheibani et al. (2019), Coates and Martin (2019), Fukas et al. (2021), Krijger et al. (2022), Mylrea and Robinson (2023), Noymanee et al. (2022), Schaschek and Engel (2023), Schmidt et al. (2022), Schuster et al. (2021), Sonntag et al. (2024), Uren and Edwards (2023), and Yams et al. (2020)	12
Organisation & Management	The structural and managerial aspects that support ethical AI implementation, including clear roles, responsibilities, and accountability. It emphasises the need for strong top management support, active stakeholder engagement, and a well-defined distribution of roles	Alsheibani et al. (2019), Coates and Martin (2019), Fukas et al. (2021), Krijger et al. (2022), Sonntag et al. (2024), and Yams et al. (2020)	6
Governance & Processes	The creation and implementation of policies and guidelines for ethical AI practices. It involves internal checks and balances, risk management, and compliance, ensuring that AI systems are developed and deployed responsibly	Coates and Martin (2019), Dotan et al. (2024), Ferreira et al. (2023), Hartikainen et al. (2023), Holmström (2022), Krijger et al. (2022), Mylrea and Robinson (2023), Schaschek and Engel (2023), Schmidt et al. (2022), and Uren and Edwards (2023)	10
Data Management	The integration of ethical considerations throughout the data science lifecycle, focusing on responsible data collection, privacy assurance, and bias mitigation in the acquisition, preparation, and management of data for AI applications	Akkiraju et al. (2020), Alsheibani et al. (2019), Coates and Martin (2019), Fukas et al. (2021), Hartikainen et al. (2023), Krijger et al. (2022), Mylrea and Robinson (2023), Noymanee et al. (2022), Schaschek and Engel (2023), Schmidt et al. (2022), Schuster et al. (2021), Sonntag et al. (2024), Uren and Edwards (2023), and Yams et al. (2020)	14

Technology	The tools, infrastructure, and workflows that support the AI lifecycle. It includes fairness assessment tools, security measures, and safeguards against risks like unauthorised access and adversarial attacks	Alsheibani et al. (2019), Coates and Martin (2019), Ferreira et al. (2023), Fukas et al. (2021), Holmström (2022), Krijger et al. (2022), Mylrea and Robinson (2023), Noymanee et al. (2022), Schaschek and Engel (2023), Schmidt et al. (2022), Schuster et al. (2021), Sonntag et al. (2024), Uren and Edwards (2023), and Yams et al. (2020)	14

Table 4.6: Dimensions of the conceptual model

4.3.1 Strategy

The first dimension identified from the literature is **Strategy**. The concept has been mentioned in seven different maturity models. Yams et al. (2020) mention that strategy addresses the ability to align and integrate AI into the broader business context and provides the "why" and "what" of organisations concerning AI activities. Fukas et al. (2021) add that the dimension of Strategy & Management, as they call it, describes the planning and formulation of objectives and strategies for the use of AI in a company regarding content, extent, temporal and spatial reference, and how the management of the firm could enable the use of AI. These descriptions are created for AI as a broader concept. However, they can also be related to Responsible AI: How does an organisation picture the responsible use of AI in the future? What changes need to be made to the organisation to reach this goal? Therefore, this dimension is crucial for understanding the Responsible AI maturity of an organisation.

The definition given to the Strategy dimension is as follows:

> *The overarching vision for how organisations will operate in the future using Artificial Intelligence and the plans on how to communicate the efforts on Responsible AI.*

4.3.2 Culture & Competences

One of the most frequently mentioned dimensions is **Culture & Competences**. Most models from the literature highlight the importance of training and the mindset of people in the organisation. The dimension's name is derived from Fukas et al. (2021), as it gives the best impression of what is included in this dimension.

Regarding training, Schmidt et al. (2022) highlight the lack of technical expertise, while Krijger et al. (2022) emphasise the training of data scientists and managers on the ethical aspects of AI. AI practitioners and Responsible AI educators also share aspirations for learning more about the social and cultural components, not just the technical angles of AI (Madaio et al., 2024). They believe this approach will help identify the social impacts of algorithmic systems early in the design process. It is, however, important to provide

employees with enough support to have potentially difficult, value-laden conversations with co-workers.

Another aspect of this dimension involves fostering an organisation-wide mindset prioritising AI ethics. A crucial aspect is enabling developers to understand that the technology they create is intertwined with ethical dimensions and that they have a vital role and responsibility to include these ethical considerations (Borenstein & Howard, 2020). It is important to have clear communication and narrative around AI ethics and to change the mentality from scepticism to enthusiasm Fukas et al. (2021).

The dimension of Culture & Competences is defined as:

> *the collective mindset, skills, and training in the context of AI ethics within an organisation. It highlights the need for continuous education and awareness of technical and ethical aspects of AI.*

4.3.3 Organisation & Management

While the **Organisation & Management** dimension is slightly related to Culture & Competences, it has distinct characteristics that justify its separation into a different category. Whereas Culture & Competences focus on people and their mindset, the Organisation & Culture dimension is more broadly oriented, encompassing roles and responsibilities in AI usage within the organisation. According to Sonntag et al. (2024) and Alsheibani et al. (2019), maturity in this dimension is achieved when roles, responsibilities, and accountability are delineated within each AI project. Moreover, a mature organisation has a clear structure that acts as a catalyst for implementing Responsible AI, ensuring its recognition. Key maturity indicators in this dimension include top management support, ecosystem participation (i.e. engagement with internal and external stakeholders, partners, and collaborators), and well-defined role distribution.

This dimension can be defined as:

> *The structural and managerial aspects to support ethical AI implementation, including clear roles, responsibilities, and accountability. It emphasises the need for strong top management support, active stakeholder engagement, and a well-defined distribution of roles.*

4.3.4 Governance & Processes

The **Governance & Processes** dimension builds upon the Organisation dimension by focusing on processes, including policies, guidelines, and governance, as noted by Schaschek and Engel (2023). Unlike other maturity models, Krijger et al. (2022) distinguish policy and governance as two distinct dimensions. However, it may be more practical to merge governance and policy into a single dimension and define them as separate indicators. Policies serve as tools for governance, with their creation being the first step, followed by implementation through governance. Governance can be seen as internal procedural ethical checks and balances in the development and deployment of AI systems (Krijger

et al., 2022). This description aligns with Dotan et al. (2024), who use the NIST definition of governance, stressing the cultivation and implementation of a risk management culture within organisations designing, deploying, evaluating, or acquiring AI systems. Coates and Martin (2019) focus on compliance, emphasising whether organisations keep records of processes such as design decisions, reasoning, and potential implications. This dimension can be defined as:

> *The creation and implementation of policies and guidelines for ethical AI practices. It involves internal checks and balances, risk management, and compliance, ensuring that AI systems are developed and deployed responsibly.*

4.3.5 Data management

Data management is arguably the most crucial dimension of this maturity model, as highlighted in numerous articles. It requires the integration of ethics throughout the various stages of the data science lifecycle within an organisation Krijger et al. (2022). This dimension involves ensuring data quality through diverse datasets, accurate labelling, and comprehensive metadata to understand data origins. Ethical considerations must be addressed during data collection, analysis, and model evaluation. The preparation, storage, and dissemination of data could impact the privacy or anonymity of subjects or introduce bias into the resulting analytics, potentially causing a data science model to operate incorrectly Saltz and Dewar (2019). Responsible data collection, privacy assurance, and bias mitigation are essential, with transparency in data handling processes building trust through ethical practices. The Data Management dimension can be defined as:

> *The integration of ethical considerations throughout the data science lifecycle, focusing on responsible data collection, privacy assurance, and bias mitigation in acquiring, preparing, and managing data for AI applications.*

4.3.6 Technology

The **Technology** dimension encompasses the tools, infrastructure, and workflows that support the AI solution lifecycle, from training and testing to deployment, monitoring, and retraining. Organisations should develop and utilise tools for assessing fairness in datasets and AI models, including bias detection tools and fairness assessment frameworks, to ensure ethical AI development. Ensuring the correct hardware and software are in place is crucial for prioritising security. Additionally, adressing risks related to the acquisition, implementation, and operation of AI systems within a broader software and technology environment is important.

The technology dimension can be defined as:

The tools, infrastructure, and workflows that support the AI lifecycle. It includes fairness assessment tools, security measures, and safeguards against risks like unauthorised access and adversarial attacks.

4.3.7 Maturity Levels

The maturity levels are defined so that they build on each other. The dimensions outlined above are largely aligned with those defined by Sonntag et al. (2024); thus, their maturity levels have also been used as a reference. The levels have the following characterisations:

- **Initial**: The organisation recognises the importance of Responsible AI but lacks formal policies or guidelines; efforts are ad hoc and reactive.
- **Experimental**: The organisation has started developing policies, processes, and training programs for Responsible AI, but practices have not yet been formalised. The focus is on understanding the requirements and laying the groundwork for more structured practices.
- **Practicing**: The organisation has defined a clear vision for Responsible AI. The first structured guidelines and processes are being implemented in the AI development process. There is a concerted effort to align AI projects with the organisation's Responsible AI vision.
- **Integrated**: Responsible AI is largely integrated across the organisation. A comprehensive Responsible AI strategy is defined, and the organisational culture supports and enables Responsible AI practices. AI projects have access to established structures and resources, ensuring that responsible development practices are consistently applied.
- **Transformed**: Responsible AI is fully embedded in the organisation's culture and operations. There is a continuous cycle of improvement and innovation in Responsible AI. A proactive approach to Responsible AI marks this stage.

4.4 Concluding insights on the literature review

The goal of this chapter was to answer the first sub-research question:

> **SRQ 1:** What (Responsible) AI maturity models are available in current academic literature?

Based on a comprehensive literature review, a conceptual maturity for assessing and improving Responsible AI was constructed. The model integrates insights from existing models, and identifies key dimensions and levels that characterise Responsible AI. It includes foundational AI elements and ethical dimensions.

A total of 22 maturity models were identified from academic literature. The majority of models focus on improving the overall AI maturity of organisations. Only a limited number of articles address Responsible AI, trustworthy AI, or AI ethics. Notably, recent articles by Krijger et al. (2022) and Dotan et al. (2024) have developed maturity models and emphasise the ethical aspects of AI.

The literature suggested various dimensions, but Strategy, Culture & Competences, Organisation & Management, Governance & Processes, Data management, and Technology were the most distinguishable ones. The maturity levels typically span from *Initial*, where recognition of Responsible AI is emerging, to *Optimised*, where Responsible AI is fully embedded in the organisation's culture and operations, with continuous improvement and innovation. This results in the first version of the Responsible AI Maturity Model (RAI-MM) including six dimensions, and five levels.

5

Delphi Study

This chapter shifts focus from the literature to the results of the empirical design of the maturity model developed through a Delphi study. The conceptual model from the previous chapter is the foundation for further refinement in this empirical phase. As outlined in Section 3.4, the process involves a three-round Delphi study, including an initial interview and two questionnaires. This chapter aims to make the maturity model more aligned with the public sector and identify missing and redundant items. The approach detailed in this chapter aligns with the *iterative maturity model development* methodology proposed by Becker et al. (2009).

5.1 Delphi round 1

5.1.1 Interview Process

The first round of the Delphi study was qualitative in nature. Although Hsu and Sandford (2007) and Skinner et al. (2015) recommend using an open-ended questionnaire, it has been decided to conduct the first round in the form of interviews for an even more explorative setup. Following the KRNW, 32 panellists were identified. Eventually, 15 panellists agreed to participate in the three-round Delphi study. The gender distribution for the longlist was balanced at 50/50, but the shortlist was more skewed towards men, with three women and 12 men participating. The distribution of panellists according to the KRNW is shown in Table 5.1.

Reference ID	Role	Date conducted	Time
P1	Industry	08-07-2024	30 minutes
P2	Industry	12-07-2024	60 minutes
P3	Academia/Industry	23-07-2024	60 minutes
P4	Government	29-08-2024	60 minutes
P5	Academia/Government	03-09-2024	60 minutes
P6	Government	04-09-2024	60 minutes
P7	Industry	05-09-2024	60 minutes
P8	Academia	05-09-2024	60 minutes
P9	Academia	05-06-2024	60 minutes

P0	Academia/Government	09-09-2024	60 minutes
P11	Academia	11-09-2024	60 minutes
P12	Industry	11-09-2024	45 minutes
P13	Industry	12-09-2024	60 minutes
P14	Academia	12-09-2024	60 minutes
P15	Industry	12-09-2024	60 minutes

Table 5.1: Records of interview Delphi study round 1

Each panellist was invited to a one-hour interview, during which they were encouraged to identify relevant dimensions and items for a Responsible AI maturity model. The interviews followed a semi-structured format, starting with more open-ended questions and concluding with a discussion of the results found in the literature. The detailed interview protocol can be found in Appendix D.

All interviews were transcribed to capture the experts' perspectives on what aspects are important in a Responsible AI maturity model. Subsequently, each interview was coded using ATLAS.ti. For the coding, a variation of the grounded theory approach was followed. Instead of coding line-by-line, the research was coded using a more modern, flexible coding approach, as described by Deterding and Waters (2018). Based on their experience with analysing interview data, Deterding and Waters (2018) suggest a coding process that better utilises Qualitative Data Analysis (QDA) technology.

The interviews were coded over three rounds, during which the codes were progressively merged into themes.

The dimensions and items identified in the literature were used as deductive codes, while the rest emerged inductively. For the coding process, the coding manual by Saldaña (2016) was also used to write valuable memos and learn more about the different types of codes that can be created.

The interview findings are largely similar to what was identified in the literature. The dimensions, in particular, do not require significant revisions based on the interviews. However, some items that could be included have emerged from the literature.

The following sections outline the interview outcomes. The panellists' statements are written down, resulting in the development of the initial empirical maturity model, which defines each item's lowest and highest maturity levels.

5.1.2 Maturity Levels, Structure and Definition of Responsible AI

Some overarching panellists' remarks should be considered before delving into the Responsible AI maturity model's dimensions and items. These comments highlight important aspects that should be incorporated into the design process of the maturity model.

Firstly, some panellists discussed the Responsible AI **maturity levels** of public organisations. According to one panellist, the lowest maturity level should represent organisations with no prior experience with AI, as the panellist encountered many such organisations during her projects (P2). Conversely, another panellist suggested that the highest maturity

level should reflect organisations that recognise AI as an integral part of their business processes, operating as responsive and optimised entities (P10).

Furthermore, panellists touched upon their **definition and understanding** of Responsible AI. One panellist outlined three perspectives of Responsible AI (P13):

- **Responsibility of the AI**: With the first perspective, the focus is on the responsibility of technology. This focus means checking for biases and evaluating the explainability and transparency of the algorithm.

- **Responsible use of AI**: This broader perspective looks at how the algorithm is used, the goals it aims to achieve, and how users interact with the system.

- **Responsible decision-making in AI**: The third perspective addresses responsible decision-making in AI development, emphasising the importance of making and documenting informed choices: "How do you decide to use or not use AI" (P10). These choices could be about how personnel is trained or about choices for required accuracy.

Most panellists see Responsible AI as the second and/or third perspective. They underscore the importance of considering ethical, legal, social, and regulatory aspects in the definition (P2, P5, P14). This perspective makes “responsible” a broader term than "ethical", with "responsible" referring to the broader societal context (P7). Responsible AI also means incorporating values and hearing stakeholders’ voices (P12). Finally, the context in which the AI is used can also affect how responsible it is. The following example was used to describe this: "using a knife in the kitchen is entirely legitimate, but stabbing someone in the ribcage is not. That is misuse" (P6). Another panellist framed it as a “do-no-harm” approach, where the technology is used in a way that does not cause harm and considers the broader impact of its use.

Additionally, several panellists underscored the importance of mechanisms to ensure AI algorithms are free from bias (P10) and produce objective results (P9). One panellist added that Responsible AI should include traceability, allowing for the justification and explanation of choices made (P11).

Finally, as one of the panellists (P6) mentioned, an interesting distinction is the different actors defined in the AI Act: providers, deployers, and affected persons. The first two should be included in the maturity model, but the third one is harder. However, the way in which the first and second actors are included should be clear to create a transparent maturity model.

Finally, one of the panellists (P2) suggested merging the *Organisation & Management* and *Governance & Processes* dimensions, as both are closely related to governance. Therefore, these dimensions will be combined into the *Governance & Processes* dimension.

5.1.3 Strategy

The combination of interviews and literature research led to the inclusion of five items in the updated version of the conceptual model. Although sustainability was mentioned in one interview (P10), it was decided to integrate it into the maturity model level descriptions rather than making it a separate item. This decision was based on the fact that no other panellists mentioned it. Moreover, the item was not identified in the literature. The items are summarised in Table 5.2 and discussed in more detail hereafter.

Item	Definition	Literature	Reference ID
Vision	A company-wide vision for the use of Responsible AI	Schuster et al. (2021) and Sonntag et al. (2024)	P1, P2, P3, P4, P6, P10, P11, P14
AI roadmap	A strategic plan that outlines the initial assessment phase to advanced implementation of AI, as well as long-term requirements	Sonntag et al. (2024)	P1, P2, P4, P6
Policy	The establishment and enforcement of guidelines, rules, and standards that govern the development, deployment, and use of AI technologies	Krijger et al. (2022) and Schaschek and Engel (2023)	P2, P4, P6, P7, P8, P10, P11, P13, P14
AI architecture	The integration of AI technologies in the IT landscape	N/A	P2, P4, P10
Investment management	Availability and management of investment capital for the implementation of AI projects	Sonntag et al. (2024)	P10, P13, P15

Table 5.2: Item definitions for Strategy dimension after Round 1

Vision

In a strategic vision, it is essential to identify where AI can be meaningfully utilised, considering both current and future capabilities (P6). Staying informed about AI developments is crucial to understanding where the organisation can be in ten years. Additionally, communicating this vision is vital (P2). Questions to consider include: "To what extent is the vision known to employees? Is it being followed? And how well does it align with other organisational strategies?" (P2).

Another panellist highlighted the importance of organisation-wide goals (P14), which should be translated into clear, concrete goals at the departmental level (P10). For example, a hypothetical goal could be “to archive all emails with AI within a year.” This task is a concrete objective that a small team can start working on and measure. Thus, while organisation-wide goals are necessary, they must also be tangible for individual professionals.

A comprehensive strategy should focus on the ICT perspective and connect with broader public domain tasks, integrating with various transition tasks (P4). For example, how AI can help urban development.

Lastly, some panellists noted that municipalities and similar public bodies often lack a coherent strategy, with many continuing long-standing practices without significant change (P3, P4). One panellist mentioned that although the Association of Dutch Municipalities published a paper on an AI strategy and vision, many municipalities have yet to implement a clear vision on AI's usefulness.

AI roadmap

An AI roadmap is closely related to the vision but serves more as a strategic outline, as the definition in Table I.6 also describes. Most panellists agreed on the importance of including a roadmap. One panellist noted that some organisations and policymakers are hesitant to adopt AI because they believe they must reach a certain maturity level first (P1). However, as outlined in the vision and echoed by several panellists, an organisation should have a clear idea of its AI goals for the next decade, including what they intend to use AI for and what they do not intend to use it for (P2, P6).

Policy

Across all panellists, policy was the most frequently mentioned item. One panellist highlighted the necessity of implementing boundary conditions, referring to them as "AI guardrails," to ensure AI usage remains within ethical, legal, and technical limits (P7). These guardrails include guidelines for AI use, addressing the current issue of widespread, unguided use of tools like ChatGPT, which poses significant risks (P13). Beyond the AI Act, organisations should develop their own criteria for acceptable AI applications (P11). Additionally, there should be guidelines on when to develop AI in-house versus when to purchase it, as well as protocols for using higher-risk algorithms (P13).

A centralised policy is needed to outline the AI applications that can and cannot be used (P13). The alignment of AI systems with organisational norms and values was also stressed, as misalignment could lead to a lack of stakeholder buy-in (P8). Moreover, government organisations have policies at various levels, from national to departmental, making it crucial for these policies to be known and aligned (P2).

Another panellist noted that policy is a broad concept and expressed interest in how it would be integrated into the maturity model (P10). A concrete example for the policy item was provided by a panellist who described how an ethical committee in a municipality developed guidelines for data and dashboard usage, emphasising transparency (P4). This panellist envisioned similar practical guides for AI applications, with another panellist also underscoring the importance of a code of conduct (P15).

Finally, another panellist mentioned that all frameworks, tools, and assessments are considered policy instruments. So, besides policy documents, it is also important to

consider policy instruments in the maturity model (P2).

AI architecture

Several panellists highlighted the importance of incorporating Enterprise Architecture, mapping AI architecture onto the broader IT architecture (P2, P4, P10). All three panellists that mentioned this item agreed that the architecture should be included in the Strategy dimension, with one noting its strong connection to policy (P2).

Investment management

Two panellists highlighted the importance of investment management in rolling out AI projects (P10, P13). It often takes over six months to secure all the necessary budget approvals. Therefore, the financial aspect of the organisation also needs to be mature to achieve Responsible AI maturity.

Another point raised was the cost of running an AI model (P15). Models with more parameters are more expensive but also more accurate than simpler ones. This dilemma raises the question of whether it is responsible to proceed if there is insufficient funding to build an accurate model that leads to less bias.

5.1.4 Culture & Competences

The second dimension that was often cited by panellists was Culture & Competences. Most of the items identified during the interviews were also corroborated by the literature. Both the literature and panellists emphasised the significance of organisational culture. However, it was decided to categorise these aspects under other items, such as training and knowledge management. Importantly, knowledge management encourages employees to share information and learn from one another, embedding it within the organisation. Discretion was the only item not identified in the literature. However, it is a concept derived from policy literature, as detailed in the item description of Table 5.3.

Item	Definition	Literature	Reference ID
Training	Continuous education for employees	Alsheibani et al. (2019), Fukas et al. (2021), Schaschek and Engel (2023), Schuster et al. (2021), and Sonntag et al. (2024)	P3, P4, P5, P6, P8, P10, P11, P13
Active management support	Leaders promote and encourage AI, and nurture an innovation and growth mindset	Alsheibani et al. (2019)	P1, P8, P9, P12
Knowledge management	Facilitate the learning and sharing of knowledge between all employees	Sonntag et al. (2024)	P2, P4, P5, P6, P9, P13, P15

Diversity	Diverse development teams	Coates and Martin (2019) and Schaschek and Engel (2023)	P2, P8
AI competences	Personal competences that employees need to possess to develop, use and improve AI technologies in an organisation	Fukas et al. (2021) and Sonntag et al. (2024)	P1, P3, P4, P6, P11, P13, P14
Ambassadors	Employees who actively promote and advocate for Responsible AI practices	Schaschek and Engel (2023)	P8, P10
Discretion	The ability of employees to make decisions and exercise judgment, even when AI systems provide recommendations or outputs	N/A	P6, P8, P9, P10, P11, P14, P15

Table 5.3: Item definitions for Culture & Competences dimension after Round 1

Training

Training within the organisation involves learning how to use AI responsibly and systematically reflecting on its use over an extended period, such as a year (P3). Simply sending an email instruction is insufficient to bring the topic to people's attention (P8). Raising employee awareness is crucial, especially since many are AI novices (P6, P10). Awareness campaigns are necessary to ensure AI is integrated into everyday tasks. While policy documents are helpful, translating them into practical instruments to create awareness within the organisation is far more effective (P4).

Expecting someone to work effectively with a new tool without proper training is unrealistic. Active training is essential to alleviate fears about using AI (P11), also known as AI demystification (P13). This concept is related to AI literacy, as described in the AI Act (P13). Employees working with AI should understand what the system does, its capabilities, and its limitations. Organisations must ensure that employees reasonably understand how AI systems work, making clear that AI is not a magical black box that solves everything but rather a set of tools that perform calculations (P13). This helps avoid situations where an AI tool is developed or experimented with, only for the organisation to be unprepared for its implementation due to a lack of support, usage, or fear.

Some other aspects of training mentioned by the panellists include training to understand the risks associated with AI (P6). For example, developers must consider what happens if somebody presses the wrong button. Additionally, training must be future-proof, ensuring it remains relevant as AI technology develops (P5). Finally, an understanding of when specific regulations, such as the General Data Protection Regulation (GDPR) and the AI Act, apply should also be included.

Active management support

One panellist emphasised the importance of making Responsible AI a top priority for management (P1). Early adopters at the top levels of the organisation are essential to driving Responsible AI initiatives forward (P8). Another panellist highlighted his experience organising sessions with stakeholders when a new AI application is being rolled out to understand the questions, worries, and context of different stakeholders. Management's active involvement and accountability in these sessions are vital (P12).

Additionally, higher management should be aware of all AI developments, ensuring that knowledge and responsibility do not remain confined to the data science team (P9). This comprehensive leadership involvement helps foster a culture of responsibility and awareness throughout the organisation. n.

Knowledge management

Panellists highlighted the importance of knowledge exchange between municipalities to facilitate mutual learning (P4). While healthy competition between departments can drive innovation, it is also important to learn from each other's mistakes (P13, P15). Organisations should be receptive to external information, sharing successes and failures to enhance collective understanding (P15). Adequate documentation is essential; without it, there is a risk that, after a year, the purpose and authorship of code may become unclear, leading to redundant efforts (P6, P15).

AI projects should be comprehensible to senior staff and non-technical teams (P9). Moreover, collaboration between technical and non-technical teams, such as data scientists and policy officers, fosters a shared understanding and language (P9, P5).

An intranet page could be valuable for addressing straightforward questions, such as legal matters or dataset usage (P5). Additionally, well-documented projects provide insights into potential pitfalls and successes, enhancing organisational learning. Organisations should consider long-term and short-term aspects, ensuring that resources are available for immediate issues (e.g., intranet page) and future planning.

Finally, organisations should also consider hosting events or encouraging employees to attend conferences and collaborate with universities to ensure the responsible use of AI (P2, P5). Formal and informal interactions, such as colloquiums, hackathons, and exchanges between organisations, can facilitate mutual learning and identify areas of strength and improvement (P5). An example of an exchange, as mentioned by one of the panellists, could be that someone from the Ministry of Defence joins the Police for six months and vice versa, allowing both parties to learn from each other and understand their respective strengths and weaknesses.

Diversity

A multidisciplinary team is essential to ensure the organisation is well-informed, well-trained, and consistently aware of handling AI applications (P8). Equally important is the diversity of knowledge backgrounds among those developing AI. In a development team, applying diversity and inclusion is crucial. This approach ensures that individuals can advocate for their own interests during the development process, thereby addressing potential biases and ensuring a more comprehensive and inclusive AI development (P2).

AI competences

Several panellists emphasised the necessity of technical competences for the responsible use of AI (P4, P6, P13, P14). A fundamental understanding of AI technologies, probabilities, statistics, big data, algorithms, and machine learning is essential (P7). Organisations must possess the knowledge to assess and support AI models, even if they do not build them themselves (P11, P14). This competence is crucial for evaluating whether the output results align with the policies or regulations intended for implementation (P11). Additionally, technical competencies include translating business questions into data-driven inquiries and ensuring that different parts of the organisation understand each other (P4). One panellist noted that outsourcing AI development without understanding its workings and testing processes reflects organisational immaturity (P13).

Beyond technical competences, some panellists discussed the broader scope of competences required (P1, P3, P13). One panellist highlighted the importance of AI literacy across the organisation, ensuring employees understand what AI is and how to use it (P13). Specific role-based competences are also necessary, including expertise in AI model development, bias detection, legal and ethical issues, organisational knowledge, and user interface design (P13).

Ambassadors

Two panellists highlighted the importance of having ambassadors for the responsible use of AI. Early adopters in key positions are essential to spread AI initiatives widely (P8). It is not sufficient to have only top management advocating for AI. Ambassadors are needed at various levels within the organisation (P10).

Discretion

Panellists highlighted the importance of discretion in the use of AI within organisations. One panellist referred to it as having a human-in-the-loop (P14). Another panellist introduced the term 'discretion' to describe how people interact with AI, noting that some organisations are biased towards mindlessly following AI recommendations (P9). This concept contrasts with the 'algorithmic colleague' approach, where professional judgement is more integrated into the organisation, as described in policy literature

(Meijer et al., 2021). A culture that rewards adherence to AI recommendations can lead to over-reliance on the model (P11).

Another panellist provided an example of a doctor using AI to diagnose cancer (P10). The critical question is whether the doctor should rely on their expertise or the AI's suggestion and whether the AI's results should be reviewed before or after making a diagnosis. This issue extends to public sector evaluations, such as those conducted by Dienst Uitvoering Onderwijs (DUO) or the Employee Insurance Agency (UWV). Similarly, suppose an AI model advises the Police to check specific cars. In that case, it is crucial to determine whether officers should follow this advice unthinkingly or use their judgment (P8).

One panellist questioned whether responsibility should lie with the technology's selection, training, and evaluation or if an additional verification step should be included before using AI outputs. For instance, should letters and fines be issued automatically, or should there be an extra verification step (P11)? If an AI model prescribes medication and it results in a patient's death, who is responsible (P15)? Another panellist emphasised the need for the ability to manually override and re-evaluate AI models, mainly when anomalies occur (P6). For example, something goes wrong if rejections suddenly increase from 100 to 2000 out of 10,000.

5.1.5 Governance & Processes

Similar to the findings in the literature, many panellists highlighted items within the Governance & Processes dimension. The overview of the items is shown in Table 5.4. Some of these items, such as governance structure, accountability, and compliance, were also corroborated by the literature, whereas the other items do not appear in existing maturity models. Supplier management is a dimension with a description tailored explicitly to the public sector, where tendering is an important process. Additionally, unlike commercial organisations, there is a high probability that AI models will need to be procured externally due to a lack of internal expertise. Citizens play a crucial role in stakeholder engagement, contrasting with commercial organisation dynamics.

Item	Definition	Literature	Reference ID
Governance structure	Clear and defined AI roles in the organisation, from leadership to team	Krijger et al. (2022), Schuster et al. (2021), and Sonntag et al. (2024)	P1, P2, P3, P4, P5, P6, P8, P9, P10, P11, P13, P14, P15
Stakeholder engagement	The inclusion of differing levels/roles of employees, external industry experts and potential users in discussion of ideas and continued development	N/A	P1, P2, P3, P5, P7, P8, P10, P12, P13, P14, P15

Accountability	The establishment of clear responsibilities and ownership for the outcomes of AI systems	Alsheibani et al. (2019) and Coates and Martin (2019)	P1, P3, P4, P5, P8, P9, P10, P11, P14, P15
Compliance	Adherence to relevant regulations, standards, and policies governing the development, deployment, and use of AI technologies to ensure ethical and legal conformity.	Schaschek and Engel (2023) and Sonntag et al. (2024)	P2, P3, P4, P5, P7, P10, P12, P13, P15
Impact assessment	Understanding of the usefulness, risks and benefits of AI	N/A	P3, P4, P6, P7, P8, P10, P11, P12, P13, P14, P15
Supplier management	The process of evaluating, selecting, and managing AI vendors and partners	N/A	P1, P2, P3, P4, P7, P11, P13, P14

Table 5.4: Item definitions for Governance & Processes dimension after Round 1

Governance structure

Several panellists highlighted the importance of a well-structured governance framework deeply embedded within the organisation. Mature organisations demonstrate flexibility and frequently adapt their structures to meet evolving needs (P15). The roles within an organisation are also evolving with the integration of AI (P11). While most public organisations maintain hierarchical structures that minimise the escalation of incidents, particularly in politically sensitive environments, it is advantageous for the responsible application of AI to involve higher-level management such as aldermen, directors, or managers (P3). Additionally, employees should be actively engaged in discussions about the use of Responsible AI and should feel empowered to voice concerns if issues arise (P9).

Furthermore, the significance of a multidisciplinary governance structure was emphasised (P8). Establishing dedicated roles for all aspects of AI development and utilisation, including technical and legal positions, is essential (P5). The governance framework should facilitate interaction and collaboration among these roles rather than allowing them to operate in isolation. Mature organisations typically have dedicated roles or teams for these responsibilities rather than treating them as ancillary tasks.

Clarity in delineating responsibilities and decision-making authority is crucial to avoid confusion and ensure effective decision-making processes (P13). Data scientists often make decisions unconsciously, but this should not always be the case. Decisions regarding using AI and assigning responsibilities should be made at appropriate levels within the organisation. Additionally, mechanisms should be established to resolve disagreements and determine decision-making authority (P5).

The importance of roles such as Data Officer, Privacy Officer, AI Officer, and ethical committees was frequently mentioned (P1, P2, P4, P6, P10, P14, P15). While some

public organisations have established these roles, there is often a lack of clarity regarding their responsibilities. Nonetheless, these roles are critical for guiding the organisation's direction in AI and ensuring AI awareness throughout the organisation (P13, P6). Panellists agreed that the specific roles required might vary depending on the organisational context, and there needs to be a balance with existing roles, with some potentially being phased out to make way for new ones (P14). Organisations must identify the roles necessary for responsible AI governance (P10).

Establishing an ethical committee or council to evaluate AI solutions from multiple perspectives was also important (P2, P6, P15). Such ethical processes enable transparent and open demonstration of the choices, their rationale, and how key public values have been safeguarded (P4).

Stakeholder engagement

Achieving stakeholder buy-in is essential (P1), and this involves engaging a diverse range of stakeholders to gain insights from various perspectives (P3, P15). Stakeholders should be engaged in a multidisciplinary setting to discuss the ethical dilemmas that may arise from the algorithms you wish to develop or apply (P14). This ensures that any blind spots are brought to light. Depending on the organisation, this could be done through a panel. According to one of the panellists, there are four types of groups that should be involved: people who will professionally work with the AI application, people who are affected by it, people who develop the AI, and those who handle policy and management (P12). Each group should share what they find important, what effects they fear, and which effects they find beneficial. Involving these groups stems from the idea of "practical wisdom": ethical knowledge resides with the people who are using it or are affected by the AI application. Engaging with these stakeholders also ensures that people are not disproportionately affected relative to the goal the organisation aims to achieve (P7). Additionally, there should be space for ongoing discussion during the actual use of the AI to ensure that all important aspects are sufficiently considered (P14).

An example of involving stakeholders is as follows (P8): "Suppose you use drones to measure crowd density in the city centre, with an AI model determining when it is too crowded. You need to inform and involve the stakeholders. Street supervisors need to know how to use it. Citizens need to know how images are processed. Businesses and shops in the city centre also have a stake. All these stakeholders are interested in the proper development of the AI model and should be able to provide feedback". There should also be effective communication, with a media strategy, within your organisation or towards society to inform stakeholders about what you are doing (P2, P5). Employees within the organisation should also feel like important stakeholders in the change (P10, P13).

Accountability

Concerning AI, accountability is a two-sided concept that needs to be well-defined (P14). It encompasses both the responsibility for the AI model itself and the broader obligations towards citizens and other stakeholders.

Firstly, it is important to identify who is responsible for each AI model in use (P15). These responsibilities must be continuously considered and explicitly defined, extending beyond mere auditing (P3). If an incident involves AI, it is essential to determine who will be held accountable (P14). When procuring AI solutions, public organisations must establish agreements with partners regarding legal responsibilities and how they will handle ethical breaches throughout the supply chain (P1). The traceability of AI model outcomes is also vital. Organisations must be able to justify their decisions and explain why they are appropriate (P11).

Moreover, accountability extends beyond individual responsibility. Political accountability is also significant, as highlighted by one panellist (P9). Public organisations should provide transparency about their AI models, including detailed information on when and how they plan to use them. Citizens should be informed about what is done with collected data (P8), and they should be informed about known issues and how they are or were mitigated (P5).

Currently, the government does not provide sufficient insight into the functioning of its algorithms (P10). For instance, the algorithm register in the Netherlands is not adequately updated. Accountability should extend beyond such registers (P9). For example, each AI project could include a publicly accessible report detailing the responsible parties (P9) and an ethical parallel process to demonstrate adherence to public values (P4).

Citizens should also be able to report issues if there is discriminatory behaviour or unfair treatment of certain groups (P5). What are the processes if such issues occur? If someone disagrees with a decision, what are the processes? Who should you contact if you find a bug in the AI system?

Compliance

Compliance involves not only adhering to relevant regulations such as the GDPR, Digital Services Act (DSA), copyright law, and the AI Act (P10, P12, P13, P15) but also actively maintaining control (P3). At a lower level of maturity, organisations focus on achieving compliance with these regulations. In contrast, a higher level of maturity indicates proactive management and anticipation of compliance requirements. For instance, if a law becomes obsolete, any advice based on that law should be removed from the system (P7). Achieving a higher maturity level also necessitates the implementation of a continuous plan-do-check-act (PDCA) cycle for ongoing improvement (P2).

Each system should undergo a simplified checklist to ensure adherence to relevant legal standards (P5). This checklist requires a thorough understanding of the applicability of various laws. Maturity also entails monitoring and regulating internally, with transparency

being crucial. Understanding the system and its data facilitates questioning and improving the model. It is essential to have a control system, which can be either internal or external.

Adhering to the organisation's ethical frameworks and regulatory compliance are vital (P4). Consensus on ethical norms and values is particularly significant in the public sector (P5).

Impact assessment

Before deploying an AI solution, it is essential to clearly define the goals and expectations for the system (P6). The concept of "purpose" is increasingly important (P3). Reflect from an ethical perspective on why you are using AI and why now (P3, P15), and create a checklist of the benefits and potential risks (ethical and privacy-related) before implementation (P10) - understanding that AI is not a universal solution and requires significant expertise (P6). Evaluate whether AI is an appropriate response to the problem or if there are alternative solutions that do not require AI (P4). An AI solution should maintain or improve the quality of work without making tasks more tedious (P7). The technology should be used for the right reasons, not merely because it is a trend (P8). When considering its use and added value, consider what might be lost when adopting AI (P10). One panellist noted, "Do not use a cannon to kill a mosquito" (P12).

An impact analysis should be comprehensive. Understanding who will be affected by the solution (P3) is necessary. Assess the impact of AI outcomes on society, the organisation, and customers, and determine whether the organisation can bear the responsibility for the impact (P11). Potential issues should be identified early, and strategies for mitigating these risks should be developed (P4, P6). Clearly define and document business processes, communication processes, and interaction patterns, including critical moments where an AI model might make decisions that impact outcomes (P14).

Decisions regarding acceptable accuracy and bias levels must be made during AI development. All decisions should be documented, including who makes them, where in the process the AI system is implemented, and how to identify when something goes wrong (P13).

Supplier management

Organisations need to draft data processing agreements with model developers (P1). In addition to the data processing agreements, thorough management of suppliers and subcontractors (P3, P4) is also needed. An organisation also needs insight into agreements with stakeholders in the supply chain. These measures should help adequately cover ethical risks (P2).

Organisations should also have precise requirements when purchasing AI models (P14). Many organisations lack clear procurement guidelines for AI (P13). Do not just purchase any black box AI solution; set requirements for explainability and responsibility (P11). Like other IT systems, you must understand what you want to achieve beyond costs

and implementation time (P7). Establish procurement requirements to know what to look for, when to purchase, and what the supplier must meet (P13). When you buy an AI solution from a commercial organisation, they may have made different technical choices, such as data storage and integration between datasets, than you would have made if you had developed the model in-house (P14).

5.1.6 Data & Information

For the Data & Information dimension, the items were primarily derived from the literature, as they were mentioned less frequently during the interviews. Data quality was the only item consistently highlighted as important from various perspectives, whereas metadata was not mentioned. The literature also references other items such as data administration, data models, and data administration (Schaschek & Engel, 2023; Sonntag et al., 2024). However, these were incorporated into other dimensions or items. For instance, data administration significantly overlapped with compliance. The items that have been included in the maturity model are shown in Table 5.5.

Item	Definition	Literature	Reference ID
Data quality	The quality of the data for training the AI models	Fukas et al. (2021), Schaschek and Engel (2023), Schuster et al. (2021), and Sonntag et al. (2024)	P2, P3, P6, P7, P9, P10, P11, P12, P13
Metadata	Metadata management system that creates and stores metadata describing important functional entities as data is processed	Sonntag et al. (2024)	N/A
Data ecosystem	A system that simplifies the management and use of data across various applications	Fukas et al. (2021) and Sonntag et al. (2024)	P11, P13, P14
Data policy	Applied structures and rules for processing the data	Sonntag et al. (2024)	P13

Table 5.5: Item definitions for Data & Information dimension after Round 1

Data quality

Data quality emerged as a frequently mentioned concern among panellists. While one panellist suggested that data quality should not pose a significant problem due to the capability of LLMs to generate synthetic data and fill gaps where necessary (P1), the majority highlighted that public organisations are lagging behind in ensuring good data quality (P3). One of the issues is the lack of consistent data quality monitoring and storage

practices, which can vary significantly across different parts of an organisation (P13). Sometimes, data is not stored correctly and only documented on paper or emails.

Several panellists discussed the garbage-in-garbage-out principle, noting that poor input data inevitably leads to poor outcomes (P2, P7, P10). One panellist stressed the need to establish conditions for the data used to train AI models (P7). Another mentioned that periodic checks on data quality are essential, preferably conducted semi-automatically or automatically. These checks should include assessments of completeness, timeliness, and the absence of bias (P3, P12).

It was noted that there are two types of relevant data: the data used to train the model and the data provided by the organisation for operational use. The latter is often more accessible, whereas obtaining high-quality training data can be particularly difficult, as most organisations do not have access to AI model training data (P11).

In risk management, selecting good training data is crucial to minimise the risks of false positives and false negatives, particularly in sensitive applications. Identifying risks and adjusting datasets is key to this process (P6).

Finally, data-centric AI was discussed, highlighting a shift from focusing solely on model refinement to ensuring data quality and reliability. For example, policing data is often very imbalanced, and simulation approaches and upsampling could help to balance out the data and thus improve data quality (P9).

Data ecosystem

Ethical and legal considerations are needed when linking data sources (P14). For AI models, there will be more data integrations (P11). Clear agreements for sharing and processing data must be established to ensure consistency and reliability (P13).

One panellist provided an example of a project combining data from different regions to train an algorithm. The data, however, had been collected in various ways and formats, which posed significant difficulties for effective integration and utilisation (P13).

Data policy

One of the panellists emphasised the importance of identifying the data owner, understanding the allowed uses of the data, and knowing whether it can be shared (P13). The other participants did not explicitly mention this item, hence the short description.

5.1.7 Technology & Tooling

The Technology & Tooling dimension had the fewest items mentioned, with only two identified. Nevertheless, both items were frequently cited in the literature and during the interviews. Security was also mentioned twice but overlapped with infrastructure, leading to the decision to merge these two items. Table 5.6 shows the definitions of the two items.

Item	Definition	Literature	Reference ID
Tooling	The functionality and quality of AI systems are constantly monitored with tests	Fukas et al. (2021), Krijger et al. (2022), and Schaschek and Engel (2023)	P2, P3, P4, P5, P8, P10, P11, P14, P15
Infrastructure	The technological components and systems that support the development, deployment, and maintenance of AI solutions	Alsheibani et al. (2019), Fukas et al. (2021), Schaschek and Engel (2023), Schuster et al. (2021), and Sonntag et al. (2024)	P1, P3, P13, P15

Table 5.6: Item definitions for Technology & Tooling dimension after Round 1

Tooling

Panellists mentioned three distinct concepts related to the tooling item. Firstly, several panellists highlighted the importance of monitoring and evaluation (P3, P10, P11, P14). It is essential to periodically assess the functioning of your AI model and examine the black box to understand the inputs and outputs (P3). Additionally, it is crucial to monitor and verify that the AI model's results are valid, reliable, relevant, and meet the organisation's expectations (P14, P11).

Secondly, testing and re-evaluation were identified as significant aspects. Continuous testing with synthetic and real data is necessary to ensure the AI model performs as expected, considering the risks associated with inaccurate test data and noise (P8). A dedicated tool to constantly monitor and improve the model's outputs when deviations are detected is also necessary (P15).

Lastly, the importance of dashboards and tools in understanding the inner workings of the black box was mentioned (P4). As mandated by the AI Act, there should be dashboards for performance and bias metrics and a clear distinction between tools used for model development and those used for subsequent monitoring (P2). Additionally, toolkits that assist both developers and non-technical personnel understand the data generated by an AI model should be used (P5).

Infrastructure

Panellists emphasised the importance of testing whether the necessary hardware is available for AI use, noting that tasks such as fine-tuning convolutional neural networks require substantial computing power (P1) and that waiting a long time for your dataset to be downloaded is impractical (P5).

Additionally, the location and management of infrastructure were identified as significant considerations. Many government organisations face bottlenecks in AI adoption due

to not being on the cloud, which is essential for managing and running models, especially with the surge of Generative AI (P1). Other panellists (P3, P15) concurred with the use of the cloud and also stressed the importance of secure storage for your data and software. One panellist mentioned that an organisation should be able to perform security checks on the infrastructure to see whether there have been any cyber-attacks and, if so, how they have been countered.

Finally, if an organisation purchases a self-learning AI model and trains it on its own data, several factors should be considered: Where will the models be hosted (P13)? How will the models be hosted? What security issues need to be addressed? How will more complex AI models be trained? What resources are required for this?

5.1.8 Outcomes of Delphi round 1

The initial round of the Delphi study resulted in the inclusion of 24 items within the maturity model. For each item, the descriptions of the lowest and highest maturity levels were defined based on input from panellists during the interviews, supplemented with insights from the literature. The maturity model developed after the first round of the Delphi study is presented in Appendix E.

5.2 Delphi round 2

In the second round of the Delphi study, panellists were asked to complete an online survey concerning the maturity model. They were asked to provide their opinions on dimensions, items, and the lowest and highest maturity level descriptions. Panellists rated each dimension and item on a five-point Likert scale, indicating the extent to which they believed a dimension or item should be included in the maturity model, ranging from "strongly disagree" to "strongly agree". If panellists selected (strongly) disagree, they were encouraged to explain their reasoning. Additionally, panellists could suggest items they felt were missing from the maturity model.

At this stage, maturity levels 2-4 have not been developed to avoid overwhelming the panellists. The objective was first to ensure that the outer levels were well-defined or to identify any changes needed. One volunteer tested the questionnaire before distribution. Details about the survey are described in Appendix F. The attrition rate between rounds 1 and 2 was 13% (n=2), with one panellist on vacation and the other unable to respond for unknown reasons.

To determine the level of consensus for a dimension or item, a method similar to that of Dragostinov et al. (2022) will be employed. Consensus on a dimension or item is reached if $\geq$ 70% or more panellists (strongly) agree with it, while no more than 15% (strongly) disagree. There is no standard practice to measure consensus, but other methods, such as the standard deviation, are used.

The following section outlines the feedback provided by the participants. Compared to round 1, the input gathered from the survey in the second round was significantly less extensive. As a result, this section is more concise than the results of round one.

5.2.1 Consensus on the items

Consensus was reached for all **dimensions**. There were no major objections to including the five identified dimensions, and none of the panellists suggested any changes. The detailed aggregated feedback for each dimension and item is provided in Appendix G.

For the **Strategy** dimension, consensus was achieved for all items but one. Over 15% of panellists disagreed with including the AI architecture item in the maturity model, so it was excluded. Key feedback indicated that panellists preferred the architecture element to be part of either the policy item or the technology dimension. Therefore, efforts have been made to integrate AI architecture into other relevant items.

In the **Culture & Competences** dimension, consensus was not reached on the it *Ambassadors* and *Diversity* items, as both failed to achieve 70% of (strongly) agree votes. However, the percentages were very close to 70%. The feedback provided, therefore, determines the next steps for these items.

One panellist suggested merging the *Ambassadors* item with *Active Management Support* into a new item called *Active Support*. Ambassadors should be present at every level of the organisation, including leadership. Therefore, combining these items was considered a logical step.

Some panellists noted that the *Diversity* item is a nuanced topic that depends on the specific task. As a result, it is not applicable in every situation and has been excluded from the next round.

For the **Governance & Processes** dimension, consensus was achieved on all items. The feedback primarily concerned the definitions of the items and their level descriptions. The only doubts come from *Supplier Management*, with some panellists mentioning that Responsible AI is not a supplier issue and that supplier management is not governed but rather an operational activity. However, the votes show that most panellists (84.7%) consider it important.

For the **Data & Information** dimension, both *Data Ecosystem* and *Metadata* scored below 70%. It appears that the description for the Data Ecosystem confused some panellists. Since its score was very close to the threshold, it will be reviewed in the next round. The same applies to the Metadata item. Some panellists suggested merging data quality and metadata. However, due to the significant difference between the two items, *Metadata* will also be reviewed in the next round to gather further opinions from other panellists.

For the **Data & Information** dimension, both *Data Ecosystem* and *Metadata* scored below 70%. The description of the Data Ecosystem confused some panellists. Since its score was very close to the threshold, it will be reviewed in the next round. The same applies to the Metadata item. Some panellists suggested merging data quality and metadata, but due

to the significant difference between the two items, Metadata will also be reviewed in the next round to gather further opinions from other panellists.

For the **Technology & Tooling** dimension, the *Infrastructure* item was also just below the maintained threshold. However, none of the panellists voted against the item and there was also no feedback on why it should be included. There, it have been decided to include it for the next round for review. There was a consensus for the *Tooling* item, even though the item descriptions should be clarified.

5.2.2 Outcomes of Delphi round 2

Based on the aggregated feedback provided in Appendix G and the voting results detailed in Appendix H, the items AI architecture, Diversity, and Ambassadors were excluded from the maturity model. Due to unclear feedback regarding Metadata and Data ecosystem, these items were carried forward to the third round for further evaluation. Additionally, Experimentation was introduced as a new item. The input from the experts was also utilised to refine and complete the descriptions for the remaining maturity levels.

5.3 Delphi round 3

The goal of the third and last round is to encourage panellists to review their scores based on the group response and see if they want to make any adjustments. Additionally, panellists can provide feedback on the maturity descriptions if they find any inaccuracies or omissions. All panellists were provided with the voting results from round 2 and the feedback that other panellists provided. The changes to this round were minor changes, indicating a convergence of opinions among the expert panel.

5.3.1 Consensus on items

For the **Strategy** dimension, consensus was reached on all items except for *Investment Management*, which did not receive more than 70% agreement from panellists. Some panellists noted that a detailed financial plan could not be made in the strategy phase. Due to its interrelatedness with these items, it was also suggested that Investment Management be added to either Vision & Goals or Supplier Management.

One panellist emphasised the importance of *AI architecture*, even though it was excluded from this round. Therefore, the updated model aims to include it in the description of one of the models.

For the **Culture & Competences** dimension, no consensus was reached for the *Discretion* and *Experimentation* items. As a result, Discretion has been excluded from the final model. Experimentation, however, was not excluded; it moved to the Technology dimension, as suggested by two panellists. The disagreement votes were related to the item's placement, as indicated by their feedback. Based on a panellist's suggestion, the remaining items were accepted, with only the *Active Support* item being renamed to *AI Adoption*.

Consensus was reached for all items of the **Governance & Processes** dimension. Although some changes were suggested for level descriptions, there was a strong agreement on all items. One panellist proposed moving the Impact Assessment item under Compliance, but since most panellists considered it relevant, no changes were made in this regard.

For the **Data & Information** dimension, consensus was reached for all items. However, some panellists suggested that the *Metadata* item should be placed under the Data Quality item. As a result, Metadata has been moved to be an indicator of Data Quality, as this was also mentioned several times in the previous round. There was an explicit agreement for the remaining items, and no changes were made.

One significant change for the **Infrastructure** dimension is the addition of Experimentation. Furthermore, the description of the *Tooling* item has been updated to include more than just fairness-related tools. Consensus was reached for both Tooling and Infrastructure, though.

5.3.2 Outcomes of Delphi round 3

The final maturity model, updated after round 3, is presented in Appendix K. The descriptions have been further refined, and investment management, discretion, and metadata items have been excluded from the final model. Although the experimentation item did not receive sufficient votes, two participants noted this was due to its placement in the wrong dimension. Therefore, it has been moved to the Technology & Tooling dimension. The final model consists of 20 items.

5.4 Concluding insights on the Delphi study

The Delphi study resulted in a maturity model consisting of five dimensions and twenty items, addressing the second research sub-question:

> **SRQ 2:** What levels, dimensions and items should be included in a Responsible AI maturity model for the public sector?

The primary findings concerning the model's composition highlight the government's responsibilities towards its citizens and companies. Unlike commercial organisations, it is more crucial for the government to involve citizens and other stakeholders, with transparency to the affected individuals and groups closely linked to this.

Furthermore, it was noted that the government will mainly use AI rather than develop it due to limited resources within the organisation. This underscores the importance of supplier management and the necessity of testing data quality, even if the government does not own the data.

6

Maturity assessment

After finalising the maturity model development, the next step in the research was to transform the model into a practical maturity assessment tool for public organisations. Developing a maturity assessment aligns with the **Conception of transfer and evaluation** and **Implementation of transfer media** steps from the model of Becker et al. (2009), discussing the development of the assessment and its components.

A maturity tool serves multiple purposes: measuring the current maturity level of specific organisational aspects, enabling stakeholders to identify strengths and areas for improvement, and determining which aspects need priority to reach higher maturity (Proenca, 2016). Therefore, the assessment is designed to show the current state, desired state, and provide a clear overview of the steps that can be taken to improve maturity for each item.

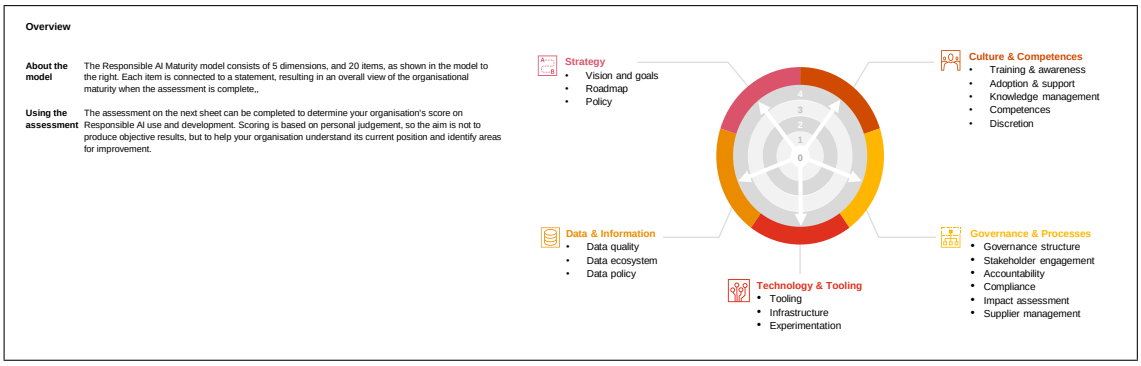


Figure 6.1: Overview page maturity assessment

The Responsible AI maturity assessment was developed in Excel, as shown above. This Excel sheet serves as a self-assessment tool and includes instructions on how to complete it. Alternatively, it can be filled out with the assistance of a consultant in a "guided self-assessment" format.

The assessment page (as shown in figure 6.2) provides an overview of the dimensions and items, along with the possibility to score an organisation on each of the items. All statements are derived from the definitions of the items established in the Delphi study. Due to time constraints, the items have not been further divided in more coarse-grained

Dimension	Item	Statement	Current state	Desired state	Explanation (optional)
Strategy	*Vision and goals*	There is a well-defined, communicated and intregated vision and goals for the responsible development and/or use of AI	3	4	
	AI Roadmap	There is a strategic plan that outlines the initial assessment phase to advanced implementation of AI, as well as long-term requirements.	2	4	
	Policy	There are guidelines, rules and standards developed that govern the responsible development and/or use of AI	2	4	
Culture & Competences	*Training and Awareness*	There is continuous awareness and training for employees, tailored to where and how AI is introduced	3	4	
	AI adoption and support	There are leaders and employees that promote and encourage the responsible development and/or use of AI and nurture an innovation and growth mindset.	4	4	
	Knowledge management	Learning and sharing of knowledge between employees about Responsible AI is facilitated	3	4	
	AI Competencies	Competences that employees need to posses to develop and/or use AI responsibly are defined, monitored and updated.	3	4	
	Discretion	Employees are encouraged to make decisions and exercise judgment on the output and recommendations of AI systems.	4	5	
Governance & Processes	*Governance structure*	There are clear and defined AI roles in the organisation, from leadership to team.	2	4	
	Stakeholder engagement	There is continuous and proactive engagement with a wide range of stakeholders, including citizens, about the use and development of AI systems.	3	4	
	Accountability	There are clear responsibilities and ownership for the outcomes of AI systems.	3	5	
	Compliance	The organisation adheres to relevant regulations, standards, and policies governing the development and/or use of AI.	3	5	
	Impact assessment	There is a clear understanding of the usefulness, risks and benefits of each AI system within the organisation.	4	4	
	Supplier management	There is a comprehensive and transparent process of evaluating, selecting, and managing AI vendors and partners.	3	4	
Data & Information	*Data quality*	The quality of data for training and using the AI system is monitored and updated.	3	5	
	Data ecosystem	There is an integrated data ecosystem that integrates and manages data, ensuring accessibility, consistency, and usability while upholding ethical principles.	1	1	
	Data policy	There are applied structures and rules for processing the data.	3	5	
Technology & Tooling	*Tooling*	The functionality and quality of AI systems are constantly monitored with tests.	3	5	
	Infrastructure	There are technological components and systems that support the development, deployment and maintainance of AI solutions responsibly.	1	1	
	Experimentation	The organisation encourages and supports experimentation with AI technologies in safe environments to drive innovation and assess value.	4	4	

Figure 6.2: Assessment page maturity assessment

statements.

To complete the maturity assessment, there are two approaches: either use the 20 statements to evaluate maturity directly, or use the extensive maturity model, as also shown in Appendix K, to identify at what level the organisation resides.

Dimension	Current	Desired
Strategy	2,3	4,0
Culture & Competences	3,4	4,2
Governance & Processes	3,0	4,3
Data & Information	2,3	3,7
Technology & Tooling	2,7	3,3

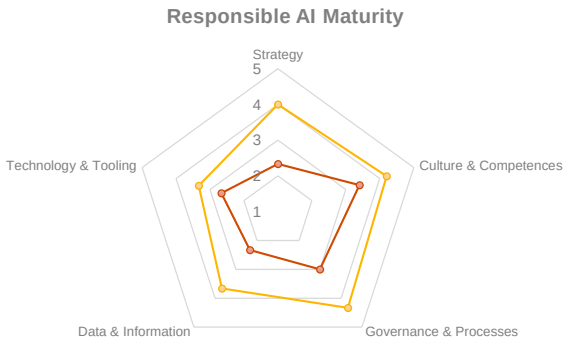


Figure 6.3: Score page maturity assessment

After filling out the maturity assessment, there is an overview that shows the current situation of the organisation and the desired situation, as shown in figure 6.3. The extensive maturity model can then be used as a reference to determine how an organisation can improve its capabilities.

7

Evaluation

As one of the final steps in this research, the iteratively developed maturity assessment was evaluated, aligning with the final step of the model by Becker et al. (2009), which involves applying the artifact in a practical setting and testing its relevance. There are different methods for evaluating a maturity model. Helgesson et al. (2011) proposes a framework consisting of three types of evaluations: Type-I, Type-II, and Type-III. A Type-I evaluation is conducted without outside experts, a Type-II evaluation involves practitioners who are experts but were not involved in the development of the maturity model, and a Type-III evaluation uses the maturity model in a practical setting.

For this research, both Type-II and Type-III evaluations are conducted:

- **Type-II evaluation**: In a one-hour workshop/session, a group of AI experts that regularly use maturity assessments in their work will be guided through the model and will collectively evaluate the model.

- **Type-III evaluation**: The maturity model will be applied in two governmental organisations to assess its practical relevance and effectiveness. It also gives a benchmark about the current maturity of some Dutch governmental organisations.

The evaluation criteria used in the evaluations are extensively discussed in Section 3.6 of the research methodology. The following two sections describe the expert evaluations and case studies in more detail.

7.1 Type-II: Expert evaluation

The maturity assessment model was evaluated with two groups: a group of academics and a group of consultants. These sessions primarily aimed to confirm whether the model contained any significant gaps or issues and to identify its academic and practical contributions, respectively. Below are the outcomes of both discussions.

7.1.1 Evaluation with academics

Six participants contributed during the one-hour session, each with varying expertise, ranging from IT project management to philosophy. The key discussion points from the session are outlined below.

The first comment raised regarding developing the maturity model through a Delphi study with experts was its potential to steer users in a specific direction, potentially leading to a *naturalistic fallacy*. This concept means that the model reflects the norm in society about responsible AI and not so much how things ought to be.

Another point raised in the discussion was who should be held accountable for ensuring responsible AI. Should the responsibility only rest on the organisation, or is a broader societal approach more suitable? Given that the model places accountability on the organisation, this is an important consideration.

Two other participants also commented on the inclusion of different groups of people in the development of the model. One participant mentioned that you need to take an as objective approach to responsible AI as possible, and that means you need to include people beyond expert opinion. Individuals who are not directly involved in AI development or use, or those with no experience with it, should also be included in the discussion about what responsible AI is.

Another participant mentioned a similar thing, mentioning that it is important to include people who will actually be using AI in the organisation. This also links to another comment made, in which one participant mentioned that responsible AI should be integrated into the broader organisation and that this required architectural planning and thorough documentation throughout the process.

Finally, one participant also mentioned that a key question is when to conduct the maturity assessment. Determining the optimal timing for the evaluation remains an open question.

7.1.2 Evaluation with consultants

Four participants were in the session with the consultants, each with varying experience with AI.

One of the group discussions focused on the maturity assessment's interface. One participant suggested that creating a user-friendly interface, possibly a web-based tool with visual elements, such as a dashboard displaying scores after completing the assessment, could enhance the maturity model's practical application. Another participant noted that whilst a polished, web-based tool might improve the user experience, there is also value in maintaining a more straightforward Excel-based format. Striking a balance between these approaches could make the model more accessible to users.

Another related discussion involved the depth and tailoring of the model's questions. The questions should be more in-depth and tailored to specific organisational contexts. Additionally, one participant emphasised that the model should help organisations identify

dependencies and interconnected items, offering clear next steps for improving their maturity. These dependencies align with the concept of a focus area maturity model.

Finally, there was a discussion about risk appetite. For example, one participant highlighted that if an organisation scores low on cultural readiness, such as having an older workforce struggling with basic technology, this should be framed as a risk to address rather than as a barrier to AI adoption. The model should provide insights into an organisation's maturity without creating resistance to adopting the technology. Another participant added that organisations should embrace AI even if they are not fully mature, as waiting too long could result in them falling behind.

7.1.3 Evaluation criteria

The participants of the expert sessions were also asked to fill out a short questionnaire to determine the ease of use, understandability, usefulness, and completeness. Ten statements could be scored on a five-point Likert scale, ranging from "strongly disagree" to "strongly agree. The questionnaire was sent out to all participants via email after the session. In total, five participants filled out the questionnaire. The results are shown in the table below. The score for understandability and usefulness was the highest, with an average of 4.3, while the score for ease of use was the lowest, with an average of 3.9.

Evaluation criteria	Questions	Score
Ease-of-use	The documentation is easy to use	4.2
	The maturity model is easy to use	3.8
	The assessment is easy to use	3.6
		3.9
Understandability	The documentation is understandable	4.6
	The maturity model is understandable	4.6
	The assessment is understandable	4.2
	The dimensions and items are understandable	3.8
		4.3
Usefulness	The maturity model is useful for conducting assessments	4.4
	The maturity model is practical for use in industry	4.2
		4.3
Completeness	The maturity model assessment criteria cover all the relevant aspects of responsible AI	4.0

Table 7.1: Components to compare maturity models

7.2 Type III: Case study evaluation

Two organisations were willing to participate in an assessment of the maturity model. For both case studies, participants received the Excel maturity tool prior to the interview and

were asked to consider it and fill it out if possible. A one-hour interview was planned to discuss the maturity assessment for each case study. The interview itself was divided into two parts:

- **Assessment**: The first part involved discussing the completed maturity assessment to understand the reasoning behind each item. This helped identify any unclear or misinterpreted items and determine whether their estimation of maturity aligned with the model. It also provided insight into how they filled out the assessment and whether they used the maturity model as a reference point.
- **Evaluation criteria**: The second part focused on discussing the evaluation criteria. Participants were asked about the ease of use, understandability, usefulness, and completeness of the model.

The following two sections discuss the outcomes of completing the model and highlight the key takeaways from the respective case studies.

7.2.1 Case: Municipality

The first case study involved a Dutch municipality. A senior advisor from the CIO office attended the meeting to evaluate the model. The model was completed during the session with the researcher and the participant.

Dimension	Item	Statement	Current state	Desired state
Strategy	*Vision and goals*	There is a well-defined, communicated and intregated vision and goals for the responsible development and/or use of AI	3	4
	AI Roadmap	There is a strategic plan that outlines the initial assessment phase to advanced implementation of AI, as well as long-term requirements.	2	4
	Policy	There are guidelines, rules and standards developed that govern the responsible development and/or use of AI	2	4
Culture & Competences	*Training and Awareness*	There is continuous awareness and training for employees, tailored to where and how AI is introduced	3	4
	AI adoption and support	There are leaders and employees that promote and encourage the responsible development and/or use of AI and nurture an innovation and growth mindset.	4	4
	Knowledge management	Learning and sharing of knowledge between employees about Responsible AI is facilitated	3	4
	AI Competencies	Competences that employees need to posses to develop and/or use AI responsibly are defined, monitored and updated.	3	4
	Discretion	Employees are encouraged to make decisions and exercise judgment on the output and recommendations of AI systems.	4	5
Governance & Processes	*Governance structure*	There are clear and defined AI roles in the organisation, from leadership to team.	2	4
	Stakeholder engagement	There is continuous and proactive engagement with a wide range of stakeholders, including citizens, about the use and development of AI systems.	3	4
	Accountability	There are clear responsibilities and ownership for the outcomes of AI systems.	3	5
	Compliance	The organisation adheres to relevant regulations, standards, and policies governing the development and/or use of AI.	3	5
	Impact assessment	There is a clear understanding of the usefulness, risks and benefits of each AI system within the organisation.	4	4
	Supplier management	There is a comprehensive and transparent process of evaluating, selecting, and managing AI vendors and partners.	3	4
Data & Information	*Data quality*	The quality of data for training and using the AI system is monitored and updated.	3	5
	Data ecosystem	There is an integrated data ecosystem that integrates and manages data, ensuring accessibility, consistency, and usability while upholding ethical principles.	1	1
	Data policy	There are applied structures and rules for processing the data.	3	5
Technology & Tooling	*Tooling*	The functionality and quality of AI systems are constantly monitored with tests.	3	5
	Infrastructure	There are technological components and systems that support the development, deployment and maintainance of AI solutions responsibly.	1	1
	Experimentation	The organisation encourages and supports experimentation with AI technologies in safe environments to drive innovation and assess value.	4	4

Figure 7.1: Case study Municipality

During the first 30 minutes of the meeting, the maturity model was filled out, accompanied by a discussion explaining the rationale behind each score. The discussion highlighted that the primary area of focus is Strategy, while Culture & Competences received the highest scores. There was some uncertainty regarding the Data & Information and Technology & Tooling dimensions, as the participant lacked sufficient information to assess topics such as the data ecosystem and infrastructure.

Regarding Strategy, the participant mentioned that there remains significant uncertainty about how the AI Act will be translated from national policy into local policy. Although the municipality is working on developing additional local policy documents, these remain abstract and lack the necessary level of detail to guide practical implementation.

On the other hand, the participant indicated that the municipality had made notable efforts to build awareness and competency around AI. Several training programs have been initiated, including awareness sessions on the AI Act, and relevant training courses have been added to the municipality's academy catalogue. During an earlier research into the use of algorithms within the organisation, the first two months were spent clarifying what an algorithm is. The participant stressed the importance of this conversation, as AI models can significantly impact both the physical and digital living environments of citizens and businesses. Understanding what constitutes an algorithm was recognised as a critical first step.

Concerning Governance & Processes, compliance was a key topic of discussion. The participant noted that the municipality follows strict GDPR rules in line with national guidelines, but the AI Act is perceived as more open-ended. For instance, the participant mentioned that transparency remains ambiguous, raising questions about how to implement it in the organisation.

The municipality did take steps to engage stakeholders, including conversations with a panel of citizens to discuss AI/algorithm-related topics. Interestingly, the participant remarked that citizens mentioned they would find it unusual if the municipality did not use algorithms and AI.

The second part of the conversation focused on discussing the evaluation criteria for the model. The participant described the model as *useful* to fill out for a group of people, as it provides a way to assess an organisation's maturity collectively. Completing the model individually might be challenging, as answering all the questions requires input from various areas of expertise. Therefore, the model scores slightly lower on the *ease of use*.

The participant found the model to be *understandable*, noting that the dimensions and items were clear and well-structured. The participant appreciated that the model is visually represented and not overly complex. This one felt less overwhelming compared to other models with as many as 80 questions.

In terms of *completeness*, the participant identified one area for improvement: the inclusion of sustainability. The participant pointed out that while AI is increasingly used for various tasks, there is little awareness of its environmental costs, such as energy and water consumption.

The key takeaways from the conversation are as follows:

- **Incorporating sustainability**: The participant identified sustainability as an important item to add to the maturity model. The participant mentioned that an average conversation with ChatGPT costs around 0.5 litres of water. People should be made aware of the environmental cost of its usage. Sustainability could be included as an item under the Culture & Competences dimension, potentially falling under training and awareness.

- **Adding follow-up questions**: The participant appreciated that the model is not overwhelming and relatively straightforward. However, the participant suggested introducing a second level to the questionnaire. If certain items or dimensions prove challenging for an organisation, the model could include additional follow-up questions to explore these areas further.

- **Difficulties with limited AI usage**: From this case study, it can be concluded that completing the model can be challenging if AI is not widely used within the organisation.

7.2.2 Case: Ministry

The second case study was conducted with a Data team within one of the Dutch Ministries. Three participants took part in the evaluation: an Enterprise Architect, a Data Scientist, and a member of the AI strategy team. Each participant completed the maturity model individually, and during the session, an aggregated version of the filled-out model was discussed.

Dimension	Item	Statement	Current state	Desired state
Strategy	*Vision and goals*	There is a well-defined, communicated and intregated vision and goals for the responsible development and/or use of AI	4	4
	AI Roadmap	There is a strategic plan that outlines the initial assessment phase to advanced implementation of AI, as well as long-term requirements.	2	5
	Policy	There are guidelines, rules and standards developed that govern the responsible development and/or use of AI	3	5
Culture & Competences	*Training and Awareness*	There is continuous awareness and training for employees, tailored to where and how AI is introduced	2	5
	AI adoption and support	There are leaders and employees that promote and encourage the responsible development and/or use of AI and nurture an innovation and growth mindset.	3	5
	Knowledge management	Learning and sharing of knowledge between employees about Responsible AI is facilitated	2	5
	AI Competencies	Competences that employees need to posses to develop and/or use AI responsibly are defined, monitored and updated.	2	5
	Discretion	Employees are encouraged to make decisions and exercise judgment on the output and recommendations of AI systems.	1	5
Governance & Processes	*Governance structure*	There are clear and defined AI roles in the organisation, from leadership to team.	2	5
	Stakeholder engagement	There is continuous and proactive engagement with a wide range of stakeholders, including citizens, about the use and development of AI systems.	3	5
	Accountability	There are clear responsibilities and ownership for the outcomes of AI systems.	2	5
	Compliance	The organisation adheres to relevant regulations, standards, and policies governing the development and/or use of AI.	3	5
	Impact assessment	There is a clear understanding of the usefulness, risks and benefits of each AI system within the organisation.	3	5
	Supplier management	There is a comprehensive and transparent process of evaluating, selecting, and managing AI vendors and partners.	2	5
Data & Information	*Data quality*	The quality of data for training and using the AI system is monitored and updated.	2	5
	Data ecosystem	There is an integrated data ecosystem that integrates and manages data, ensuring accessibility, consistency, and usability while upholding ethical principles.	2	5
	Data policy	There are applied structures and rules for processing the data.	4	5
Technology & Tooling	*Tooling*	The functionality and quality of AI systems are constantly monitored with tests.	3	5
	Infrastructure	There are technological components and systems that support the development, deployment and maintainance of AI solutions responsibly.	3	5
	Experimentation	The organisation encourages and supports experimentation with AI technologies in safe environments to drive innovation and assess value.	5	5

Figure 7.2: Case study Dutch Ministry

The first half of the interview focused on understanding how the model was completed and identifying significant areas for organisational improvement. The primary focus areas identified were strategy and governance, particularly the roadmap and accountability items. Although the organisation has an AI strategy, it has not yet been translated into a concrete policy. Regarding accountability, the data team primarily develops AI models, but there are no clear agreements on responsibility when these models are handed over to other teams. These insights were consistent with the self-assessment scores, indicating the model's effectiveness.

The second half of the interview discussed the evaluation criteria mentioned earlier. Due to time constraints, not all criteria were covered in detail. Participants generally found the model to be *complete*. One participant noted that while additional items could always be added, the current model sufficiently addresses responsible AI. Another participant appreciated the model's alignment with legal requirements, such as AI literacy, which enhances its objectivity, even calling it a "no-brainer".

Regarding *understandability*, participants found the full maturity model, along with the level descriptions, to be very useful as a reference guide, providing clear direction. However, in terms of *ease of use*, there was some confusion about whether the model should be filled out at the team level or organisation-wide. Some items seemed more applicable to team-level assessments, while others were more relevant to the organisation as a whole.

Finally, there were some key takeaways on how to improve the model:

- **Distinction between foundational and responsible AI**: Participants noted that the model does not always clearly differentiate between "AI" and "Responsible AI." Although it was a deliberate choice to include both foundational and responsible AI components, specifying the goal of each item more clearly could enhance the model's clarity.
- **Iterative completion**: Participants highlighted the importance of having different individuals complete the assessment independently, followed by a discussion to reconcile discrepancies. This approach could improve the reliability of the self-assessment.
- **Distinction between team and organisation levels**: Participants found it challenging to complete the model due to uncertainty about the perspective they should adopt. Providing better documentation or scoping guidelines in advance would be beneficial.

7.3 Concluding insights on the evaluation

The case studies demonstrate that the model is adequate for organisations to evaluate their responsible AI capabilities. Both sets of participants found the model valuable and suggested that involving more people within the organisation in completing it would foster a dialogue on enhancing these capabilities. As a result, the evaluations positively address the third research sub-question:

> **SRQ 2:** How does the Responsible AI maturity model hold up in practice?

As highlighted in the expert evaluations and case studies, the model could be improved by incorporating more detailed questions to better guide organisations in strengthening their responsible AI capabilities. For organisations struggling with a specific dimension or item, follow-up questions could help identify the most effective ways to improve.

8

Discussion

Throughout the research, various design choices were made regarding developing the maturity model, data collection, and data analysis. This chapter aims to reflect on these choices and examine their impact on the resulting maturity model.

8.1 Maturity Model reflection

The development of the Responsible AI Maturity Model (RAI-MM) represents a new contribution to the public sector's ability to navigate the ethical and technological complexities of AI. Design choices in the development process impacted the final outcome of this thesis and looking at the maturity model, the following reflective questions can be considered:

- Is the current choice of the maturity model structure justified over other types like CMMI or focus area models?
- How relevant will the model remain as AI technologies and ethical challenges evolve?
- Are the pathways between maturity levels clearly articulated in the model?

8.1.1 Choice of the model structure

The outcome of this study is a hybrid model, combining a *Maturity Grid* with a *Likert-scale questionnaire*. Following the classification described in Table 7.1, it was also possible to develop a structured model, often referred to as a CMM-like model. The advantage of a CMM-like model is that it has a more formal architecture and is likely to be more complete. However, as we have seen in the validation, participants appreciated the simplicity of the hybrid model and the ability to perform self-assessments. Therefore, the decision to use a hybrid model seems to be the right choice for a topic like Responsible AI. Additionally, one should consider the downsides of a more formal maturity model, such as the lack of flexibility mentioned in relation to ethical frameworks. This thesis has already shown that

balancing ethical principles with practical guidelines is a difficult task, and developing a CMM-like model further complicates this task.

Another structural choice for the maturity model was between a fixed-level model and a focus area model. During the empirical validation, one participant emphasised the value of having a benchmark to compare with other organisations, something fixed-level models are well-suited to provide.

However, the fixed-level model has its limitations. It does not offer a plan for incremental improvement. For instance, achieving level 4 in data quality first might be more beneficial than reaching level 2 in the data ecosystem. These interdependencies are not accounted for in the current model but could be important for effectively guiding progress.

In the context of Responsible AI, prioritising specific items is particularly challenging, as noted above. Consequently, the model also does not assign weights to items, as all are considered critical for ensuring the ethical use of AI. For future research, it might be worth exploring the concept of a focus area maturity model. Such a model could help identify whether there is a logical sequence to follow when advancing towards higher levels of maturity. This approach could be especially helpful for foundational AI items.

8.1.2 Future relevancy of the model

The maturity model has been designed with future AI technologies in mind. It aims to provide a comprehensive overview of an organisation's current maturity, regardless of the AI technology or system in use. Initially, the model focused on generative AI, which was a trending topic among public organisations in 2024. During the Delphi study, experts frequently referenced generative AI examples.

However, there has been a shift towards a new wave of AI, known as Agentic AI. These systems adaptively pursue complex goals using reasoning with limited direct supervision (Shavit et al., 2023, p. 2). Some of the key ethical challenges associated with Agentic AI include the "human-in-the-loop" requirement and increased labour displacement. The RAI-MM's foundation in broad ethical principles, such as transparency and accountability, ensures its applicability to Agentic AI systems as well.

I believe that the model's dimensions, such as governance and culture & competences, are applicable for addressing the risks posed by Agentic AI. The primary difference will be the emphasis placed on certain aspects. For instance, tooling and accountability might become more significant, as some decisions are too critical to delegate to agents. One area for improvement in the maturity model is the addition of indicators for each item that are also more closely related to Agentic AI.

8.1.3 Pathways between maturity levels

For each item, a maturity path has been created with level descriptions derived from literature and interviews, and updated through questionnaires. Although the empirical feedback was positive and the level distinctions were considered clear, a more explicit

separation of levels is needed. This could be achieved by using bullet points or clearer indicators. Initially, the goal was to further split the descriptions, but it proved more challenging than anticipated.

To further develop the maturity model, it is essential to clarify these distinctions to create a logical roadmap. The current model provides guidelines through the item maturity descriptions, but a clearer separation is necessary. Including interdependencies, as seen in the focus area maturity model, could help understand the sequence of steps, even though it may be challenging to incorporate this.

8.2 Methodology reflection

The procedural model of Becker et al. (2009) was used as the scientific basis for this thesis. It ensured the systematic development and evaluation of the maturity model and proved to be a useful method to structure the thesis. For future research, it would be useful to have a model that goes into even more detail and describes, for example, the theoretical and empirical iterations that are recommended. In this thesis, the existing maturity models formed the basis of the conceptual model, but expanding the literature review to include other papers discussing capabilities of Responsible AI could have been advantageous.

This thesis also used a combination of empirical methods to develop the final maturity model, in contrast to most maturity models found in the literature. Those models were largely conceptual and not empirically tested. When empirical testing was conducted in other models, the empirical part was often limited to a few interviews or case studies.

The main research method used in this thesis was the Delphi study, which provided valuable insights into the important dimensions and items for the maturity model. Future research could benefit from including focus groups, as it would stimulate group discussions, but this was beyond the scope of this thesis. Additionally, getting a sufficient number of experts was already a challenge, so a focus group was also not a realistic method for now.

A significant challenge during the Delphi study was maintaining panellist engagement. While the interviews in the first round provided substantial input, participation weakened during the second and third rounds. This thesis may have benefited from using the Delphi method to determine relevant dimensions and items, with the descriptions for all items discussed in a format like a focus group. This approach could have kept panellists more engaged, as providing feedback on item inclusion and exclusion requires less time commitment. Designing surveys that can be completed within 15 minutes is recommended to enhance engagement further.

For the validation, case studies and a workshop were used to assess the practical relevance of the model. However, it was not possible to validate whether the items and dimensions were mutually exclusive. A survey combined with Structural Equation Modelling would be an interesting avenue to explore.

8.3 Practical relevance

This thesis equips organisations with a practical framework to assess and improve their responsible AI practices. With the recent introduction of the AI Act, public organisations are struggling to implement changes to comply with the regulation and to reshape their operations to use AI responsibly. The practical contributions of this thesis are as follows:

1. **Improving organisational readiness**: The practical assessment tool derived from the maturity model allows organisations to identify their current maturity, desired maturity, and steps to advance their responsible AI capabilities. The case studies in the validation phase of this research also demonstrate the model's applicability and effectiveness in real-world settings.

2. **Facilitating compliance with AI regulation**: The AI Act poses challenges for organisations to comply with the law, and the RAI-MM serves as a practical tool to guide them in understanding and meeting these legal requirements. Many organisations are, for example, not aware of the importance of AI literacy, or who should be accountable for a developed AI tool.

3. **Enhancing citizen trust**: Many public sector organisations directly interact with citizens, and especially after recent scandals in the Netherlands with algorithms (such as the allowance scandal), the trust among citizens is fragile. The RAI-MM addresses citizen-centric concerns such as accountability and stakeholder engagement, ensuring that new AI tools are developed and used in consultation with citizens.

8.4 Academic relevance

This research offers academic contributions in various areas, including increasing the understanding of responsible AI, outlining methodologies for developing a maturity model, and presenting the maturity model as a key outcome of the study. The primary academic contributions are as follows:

1. **Advancing the Understanding of Responsible AI**: This thesis contributes to the current literature on Responsible AI by identifying its most important components and highlighting what experts believe constitutes Responsible AI. By incorporating dimensions such as governance & processes, and items such as stakeholder engagement and impact assessments, it provides a comprehensive framework that enriches the theoretical understanding of Responsible AI. This lays a strong foundation for future studies to further explore and refine these dimensions and items.

2. **Bridging the Gap Between Ethics and Practice**: The study offers valuable insights into the ethical principles considered as important for Responsible AI in the public

sector. Moreover, it operationalises these principles by translating them into capabilities, addressing an often mentioned academic challenge in bridging high-level ethical guidelines with real-world implementation.

3. **Development of a Tailored Maturity Model**: The Responsible AI Maturity Model (RAI-MM) developed in this thesis is the first explicitly designed for the public sector, filling a critical gap in the maturity model literature. By addressing sector-specific challenges such as citizen trust, transparency, and public accountability, this thesis introduces new items and dimensions that extend maturity model theories into previously underexplored domains.

4. **Incorporation of the AI Act**: This study demonstrates how the regulatory framework of the AI Act can be integrated into a practical framework. By aligning the RAI-MM with regulatory requirements, the thesis provides a roadmap for public organisations to navigate the complexities of compliance in an evolving legal landscape.

5. **Synthesis of Existing Maturity Models**: Building on the work of Sadiq et al. (2021), Reichl and Rudolf (2023) and Akbarighatar (2022), this thesis evaluates all existing maturity models in academic literature. It incorporates 14 newly developed models since previous reviews, thereby ensuring a cumulative and up-to-date synthesis of the field. This approach not only strengthens the academic rigour of the study but also provides a valuable resource for researchers working on maturity models in AI.

8.5 Limitations

The development of the RAI-MM involved a combination of methods, including a literature review, a Delphi study, expert sessions, and case studies. However, each of these methods has certain limitations, either inherent to the method itself or arising from biases or misinterpretations during the process.

8.5.1 Literature review

The literature review specifically focused on existing (responsible) AI maturity models, while excluding literature that addressed the broader use of AI in the public sector. Including such literature could have further enriched the study, potentially resulting in a more comprehensive list of dimensions and items critical to the maturity model. Furthermore, incorporating ancillary literature could have enhanced the descriptions and roadmap for improving responsible AI within organisations. This would result in a model offering greater depth compared to the current version.

8.5.2 Delphi study

The outcome of the literature review also presented a limitation for the Delphi study. Participants were provided with a list of dimensions and items to assess their relevance for

inclusion in the maturity model in the second round of the Delphi study. However, had the literature review been more extensive, the list might have been more comprehensive, potentially leading to a more exhaustive maturity model. For instance, sustainability was highlighted during the validation phase as an important element of responsible AI. If it had been included in the Delphi study, it would have allowed for greater certainty regarding its importance as perceived by the experts.

Another limitation of the Delphi study is that is very prone to variability. An identical setup of the thesis may result in a different outcome, despite the efforts to combine it with literature and validate it further through case studies.

Thirdly, an obvious limitation of the study is the composition of the panel. While efforts have been made to create a diverse and representative expert panel with the KRNW, the selection process may have been biased. One clear limitation is the skewness of gender among the panel experts. The invitation was sent out to an even number of female and male prospects, but the final panel only ended up with three women out of a total of 15 participants.

Finally, subjectivity in evaluating the results from Delphi studies could have led to biases affecting the maturity model. Attempts were made to make the interview and questionnaire evaluation as objective as possible, but the fact is that only one researcher was involved in the entire process.

8.5.3 Validation: case studies and expert sessions

The validation phase faced two main limitations: a small sample size and context-specific findings.

Both the case studies and the expert session involved only two organisations/groups, restricting the ability to generalise the findings to a larger population or to other types of organisations. One of the organisations was a municipality, which encounters different challenges compared to larger governmental bodies. Additionally, had health organisations been included, they might have demonstrated more hesitancy towards adopting AI across the organisation (due to concerns over patient data) and might see less value in the maturity model. Consequently, the results of the case studies are closely tied to the specific contexts in which they were conducted.

A further limitation of the case studies was the minimal number of participants from each organisation involved in the evaluation session and maturity assessment. For instance, if 15 individuals from a single organisation had completed the maturity assessment, this would have provided a more comprehensive overview of the model's value and its ability to accurately measure maturity across different roles within the organisation.

8.5.4 General

Finally, there are limitations that are not specific to any single method but apply more broadly to the study.

Firstly, the generalisability of the results to countries outside the Netherlands remains unclear. The Delphi study participants and the case studies all involved participants based in the Netherlands. Since the AI Act was frequently referenced, and elements such as AI literacy were incorporated into the maturity model, it could be argued that the model has relevance for other EU countries due to shared regulatory frameworks. Nonetheless, the broad items included from existing maturity models may ensure its applicability to countries beyond the EU as well.

Another general limitation is the single-researcher bias. The involvement of only one researcher in conducting interviews and evaluating results may have unintentionally introduced bias into the study. Efforts were made to mitigate this risk by carefully documenting the process and adhering to predefined research guidelines.

8.6 Future work

Building upon the findings of this thesis, there are several future endeavours to further refine, validate, and expand the Responsible AI Maturity Model. The goal of these future efforts would be to improve the model's validity and applicability across different contexts:

1. **Extended Literature Review:** Broaden the literature base to include insights from additional fields, such as public administration literature on AI, which were excluded from this thesis due to time constraints. Incorporating this perspective may uncover additional dimensions and items, strengthening and expanding the maturity model.

2. **Quantitative Assessment:** Conduct a quantitative assessment to evaluate the mutual exclusiveness of the model's dimensions and items. This validation step would help to refine the model further. Structural Equation Modelling would be a suitable method for this purpose.

3. **Cross-Cultural Validation:** To improve the generalisability of the model, conduct evaluations with organisations and individuals from countries outside the Netherlands, focusing on other European countries first. This cross-cultural validation would provide deeper insights into the model's applicability across different cultural and organisational landscapes.

4. **Tool Development:** Create a web-based tool to facilitate the practical application of the maturity model. This would enable organisations to measure their maturity efficiently, particularly when multiple stakeholders are involved in filling out to the assessment. Moreover, such a tool could offer instant, tailored recommendations to help organisations improve their capabilities.

5. **Benchmarking:** Incorporate a comparative component into the maturity model. While the current model is primarily a prescriptive tool, adding a comparative aspect would allow organisations to benchmark against peers, identifying relative

strengths and weaknesses. Initially, the focus should be on larger governmental organisations, as they are more likely to use AI extensively.

6. **Roadmap Development:** Develop a structured roadmap for organisations. Two key approaches could include:

   - Creating a focus area model that highlights the interdependencies between items, offering a sequential plan for achieving maturity.
   - Developing a separate roadmap to provide a clear and actionable pathway for progression.

8.7 Ethical considerations

While this thesis is primarily framed within the IS context, it also carries significant philosophical implications when viewed through the lens of AI ethics.

8.7.1 Normative ethics

Firstly, there are different schools of thought that are of particular interest for artificial moral reasoning: consequentialism, deontology, and virtue ethics. The deontological approach is the most prevalent in the field of AI ethics, emphasising that AI systems should follow universal moral laws. Practically, this results in a tick-box exercise.

Since the outcome of this thesis is a prescriptive model, there is the risk of the model becoming overly prescriptive, potentially stifling innovation and flexibility. It could also lead to ethics washing, where organisations believe they are using AI responsibly simply by ticking all the boxes or, in the case of the maturity model, by reaching level 5 of maturity.

As discussed earlier in this thesis, Hagendorff (2020) has argued that the deontological approach should be complemented with virtue ethics. This would shift the model from a checklist to a project of self-responsibility, uncovering blind spots and promoting autonomy. The maturity model should encourage organisations to go beyond mere compliance and develop a deep commitment to ethical practices, which requires continuous reflection and ethical consideration.

It could be argued that the maturity model addresses the risks of a deontological tick-box approach by including elements that allow virtue ethics to flourish. Training and awareness, knowledge management, and impact assessments are all items that keep organisations mindful of how they are using AI. The higher maturity levels of the model also incorporate a reflective aspect.

Overall, it would be interesting to further discuss whether a maturity model is indeed the best method to help organisations use AI responsibly, and whether there are levels above the "transformative" level of the current maturity model.

8.7.2 Naturalistic Fallacy

Secondly, there is a possibility that some of the dimensions and items included in the maturity model merely reflect current norms, a concept known as the naturalistic fallacy. The experts involved in developing the maturity model draw from their own experiences and contexts. Consequently, there is a risk that the items and descriptions are simply a reflection of what is currently considered the norm. A good example is the AI Act, which many experts used as a reference point. However, it is debatable whether the AI Act is genuinely related to responsible AI and whether it should serve as a reference for the maturity model. While referencing existing norms makes the model more practically useful for organizations, it may not truly represent what is "good" in terms of responsible AI. I firmly believe that including references to existing norms has made the model more practical, and bridged the gap from high-level principles, but it also means compromising on the pursuit of the most virtuous perspective of responsible AI.

For future research, it would be beneficial to involve a broader range of participants in the discussions. This should include not only experts in (responsible) AI but also individuals who are affected by AI and may not fully understand its workings. Their input could foster a more reflective and multidisciplinary dialogue, resulting in a model that better aligns with ethical values.

8.7.3 AI for Good

Finally, Responsible AI does not exist in isolation but is part of a broader socio-technical system. This implies that AI systems should not only be developed and used ethically but also for purposes that genuinely benefit society. Reflecting on the maturity model, we could question whether we are truly fostering development and use for a good cause. An organisation might score low on the impact assessment yet high on other criteria, leading to a high maturity rating. However, this could mean that the organisation is creating AI that is not genuinely meaningful or beneficial.

The inherent limitation of the current model is its non-sequential nature. Ideally, determining the strategy and conducting impact assessments should be the foundational steps in the development and deployment of AI. A sequential approach ensures that the ethical implications and societal impacts are considered from the start. The absence of this structured progression in the model represents a significant weakness. To address this, future iterations of the model should incorporate a clearer, more sequential growth path, emphasising the importance of initial strategic planning and impact assessment in the journey towards responsible AI.

9 Conclusion

As Artificial Intelligence becomes more embedded in decision-making processes, public organisations are uniquely positioned due to their direct interactions with citizens and businesses. These interactions come with significant ethical responsibilities, and maintaining public trust is a vital part of their role.

However, a gap exists between high-level ethical principles and ethical AI practices, leaving public organisations uncertain about how to use AI and assess their current practices responsibly. This research sought to address this gap by answering the following question:

> *How can a maturity model be developed and evaluated for public organisations to measure and guide their Responsible AI capabilities?*

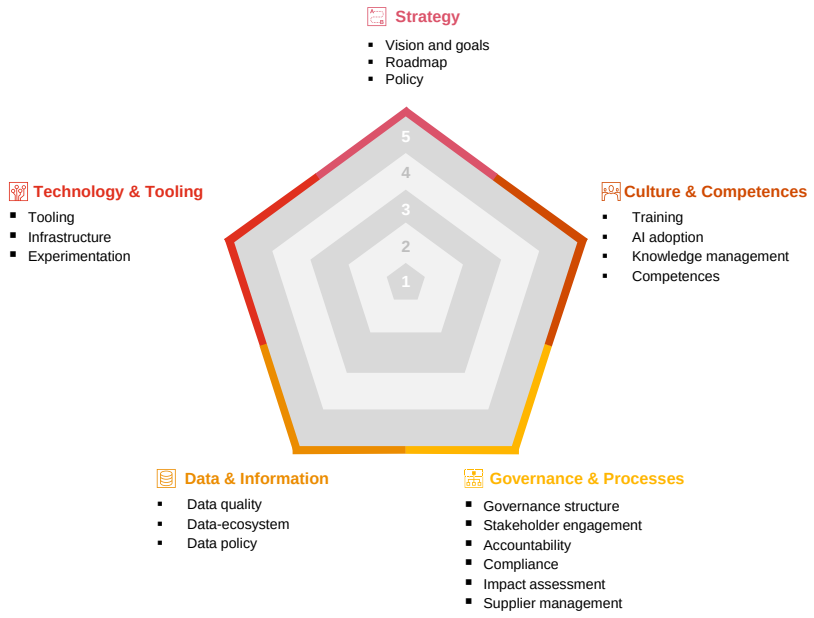


Figure 9.1: Visualisation maturity model

The findings of this thesis underscore the relevance of a tailored Responsible AI Maturity Model (RAI-MM) for public sector organisations. *Developed* through a systematic literature review and a three-round Delphi study, the model highlights that achieving responsible AI maturity is not merely a technical challenge but a multidimensional endeavour, as illustrated in figure 9.1. The key dimensions identified include Strategy, Culture & Competences, Governance & Processes, Data & Information, and Technology & Tooling, each with specific items to *measure* maturity. Organisations can use these items for self-assessment or guided discussions, with the complete maturity model offering actionable *guidance* to facilitate growth towards greater maturity.

The RAI-MM presented in this research goes beyond existing models by addressing the specific challenges and responsibilities of the public sector, such as transparency towards citizens and accountability for AI models. The RAI-MM fosters not only compliance with the AI Act but also the proactive implementation of responsible AI.

The empirical *evaluation* of this model through case studies and expert interviews shows the suitability of the maturity assessment as a descriptive and prescriptive tool, providing recommendations for further development. As AI continues to evolve, the dimensions and items established in this research will remain foundational.

Bibliography

Akbarighatar, P. (2022). Maturity and readiness models for responsible artificial intelligence (rai): A systematic literature review.

Akdil, K. Y., Ustundag, A., & Cevikcan, E. (2018). Maturity and readiness model for industry 4.0 strategy. In A. Ustundag & E. Cevikcan (Eds.), *Industry 4.0: Managing the digital transformation* (pp. 61–94). Springer International Publishing. https://doi.org/10.1007/978-3-319-57870-5_4

Akkiraju, R., Sinha, V., Xu, A., Mahmud, J., Gundecha, P., Liu, Z., Liu, X., & Schumacher, J. (2020). Characterizing machine learning processes: A maturity framework. In *Business process management* (pp. 17–31). Springer International Publishing. https://doi.org/10.1007/978-3-030-58666-9_2

Alarabiat, A., & Ramos, I. (2019). The delphi method in information systems research (2004-2017). *Electronic Journal of Business Research Methods*, *17*(2). https://doi.org/10.34190/jbrm.17.2.04

Alsheibani, S. A., Cheung, Y. P., & Messom, C. H. (2019). Towards an artificial intelligence maturity model: From science fiction to business facts. *Pacific Asia Conference on Information Systems*. https://api.semanticscholar.org/CorpusID:211165308

Anagnostou, M., Karvounidou, O., Katritzidaki, C., Kechagia, C., Melidou, K., Mpeza, E., Konstantinidis, I., Kapantai, E., Berberidis, C., Magnisalis, I., & Peristeras, V. (2022). Characteristics and challenges in the industries towards responsible ai: A systematic literature review. *Ethics and Information Technology*, *24*(3). https://doi.org/10.1007/s10676-022-09634-1

Andersen, E. S., & Jessen, S. A. (2003). Project maturity in organisations. *International Journal of Project Management*, *21*(6), 457–461. https://doi.org/10.1016/s0263-7863(02)00088-1

Andersen, K. V., & Henriksen, H. Z. (2006). E-government maturity models: Extension of the layne and lee model. *Government Information Quarterly*, *23*(2), 236–248. https://doi.org/https://doi.org/10.1016/j.giq.2005.11.008

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2021). Sociotechnical envelopment of artificial intelligence: An approach to organizational

deployment of inscrutable artificial intelligence systems. *Journal of the Association for Information Systems*, *22*(2), 325–352. https://doi.org/10.17705/1jais.00664

Bandi, A., Adapa, P. V. S. R., & Kuchi, Y. E. V. P. K. (2023). The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, *15*(8). https://doi.org/10.3390/fi15080260

Becker, J., Knackstedt, R., & Pöppelbuß, J. (2009). Developing maturity models for it management. *Business Information Systems Engineering*, *1*(3), 213–222. https://doi.org/10.1007/s12599-009-0044-5

Bibby, L., & Dehe, B. (2018). Defining and assessing industry 4.0 maturity levels – case of the defence sector [doi: 10.1080/09537287.2018.1503355]. *Production Planning Control*, *29*(12), 1030–1043. https://doi.org/10.1080/09537287.2018.1503355

Bley, K., Schön, H., & Strahringer, S. (2020). Overcoming the ivory tower: A meta model for staged maturity models. In M. Hattingh, M. Matthee, H. Smuts, I. Pappas, Y. K. Dwivedi, & M. Mäntymäki (Eds.), *Responsible design, implementation and use of information and communication technology* (pp. 337–349). Springer International Publishing.

Borenstein, J., & Howard, A. (2020). Emerging challenges in ai and the need for ai ethics education. *AI and Ethics*, *1*(1), 61–65. https://doi.org/10.1007/s43681-020-00002-7

Brown, J. (2018). *Advanced research methods for applied psychology: Design, analysis and reporting*. Routledge. https://doi.org/10.4324/9781315517971

Cho, S., Kim, I., Kim, J., Woo, H., & Shin, W. (2023). A maturity model for trustworthy ai software development. *Applied Sciences*, *13*, 4771. https://doi.org/10.3390/app13084771

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023, June). The economic potential of generative ai: The next productivity frontier: Mckinsey. https://www.mckinsey.com/featured-insights/mckinsey-live/webinars/the-economic-potential-of-generative-ai-the-next-productivity-frontier

Coates, D. L., & Martin, A. (2019). An instrument to evaluate the maturity of bias governance capability in artificial intelligence projects. *IBM J. Res. Dev.*, *63*(4/5), 7:1–7:15. https://doi.org/10.1147/JRD.2019.2915062

Cognet, B., Pernot, J.-P., Rivest, L., & Danjou, C. (2023). Systematic comparison of digital maturity assessment models [doi: 10.1080/21681015.2023.2242340]. *Journal of Industrial and Production Engineering*, *40*(7), 519–537. https://doi.org/10.1080/21681015.2023.2242340

Colli, M., Berger, U., Bockholt, M., Madsen, O., Møller, C., & Wæhrens, B. V. (2019). A maturity assessment approach for conceiving context-specific roadmaps in the industry 4.0 era. *Annual Reviews in Control*, *48*, 165–177. https://doi.org/https://doi.org/10.1016/j.arcontrol.2019.06.001

Correia, E., Carvalho, H., Azevedo, S. G., & Govindan, K. (2017). Maturity models in supply chain sustainability: A systematic literature review. *Sustainability*, *9*(1). https://doi.org/10.3390/su9010064

Cox, W. E. (1967). Product life cycles as marketing models. *The Journal of Business*, *40*(4), 375–384. http://www.jstor.org/stable/2351620

Dalkey, N. (1969). An experimental study of group opinion. *Futures*, *1*(5), 408–426. https://doi.org/10.1016/s0016-3287(69)80025-x

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319. https://doi.org/10.2307/249008

de Bruin, T., Freeze, R., Kulkarni, U., & Rosemann, M. (2005). Understanding the main phases of developing a maturity assessment model. *Australasian Conference on Information Systems*.

de Bruin, T., & Rosemann, M. (2005). Towards a business process management maturity model. *Proceedings of the 13th European Conference on Information Systems*, *521–532*.

de Bruin, T., & Rosemann, M. (2007). Using the delphi technique to identify bpm capability areas. *ACIS 2007 Proceedings - 18th Australasian Conference on Information Systems*.

Deterding, N. M., & Waters, M. C. (2018). Flexible coding of in-depth interviews: A twenty-first-century approach. *Sociological Methods amp; Research*, *50*(2), 708–739. https://doi.org/10.1177/0049124118799377

Dignum, V. (2019). *Responsible artificial intelligence* (1st ed.). Springer Nature.

Donohoe, H. M., & Needham, R. D. (2008). Moving best practice forward: Delphi characteristics, advantages, potential problems, and solutions. *International Journal of Tourism Research*, *11*(5), 415–437. https://doi.org/10.1002/jtr.709

Dotan, R., Blili-Hamelin, B., Madhavan, R., Matthews, J., & Scarpino, J. (2024). Evolving ai risk management: A maturity model based on the nist ai risk management framework. https://doi.org/10.48550/ARXIV.2401.15229

Dragostinov, Y., Harðardóttir, D., McKenna, P. E., Robb, D. A., Nesset, B., Ahmad, M. I., Romeo, M., Lim, M. Y., Yu, C., Jang, Y., Diab, M., Cangelosi, A., Demiris, Y., Hastie, H., & Rajendran, G. (2022). Preliminary psychometric scale development using the mixed methods delphi technique. *Methods in Psychology*, *7*. https://doi.org/10.1016/j.metip.2022.100103

Ellefsen, A. P. T., Oleśków-Szłapka, J., Pawłowski, G., & Toboła, A. (2019). *Logforum*, *15*(3), 363–376. https://doi.org/10.17270/j.log.2019.354

European Parliament. (2024, March 13). Artificial intelligence act [European parliament legislative resolution of 13 march 2024 on the proposal for a regulation of the european parliament and of the council on laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (com(2021)0206 – c9-0146/2021 – 2021/0106(cod))]. Retrieved 2024-05-07, from https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

Ferreira, R. M. F. D., Grilo, A., & Maia, M. J. (2023). A maturity model for industries and organizations of all types to adopt responsible ai—preliminary results. In *Lecture notes in computer science* (pp. 67–78). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-49008-8_6

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Fraser, P., Moultrie, J., & Gregory, M. (2003). The use of maturity models/grids as a tool in assessing product development capability. *IEEE International Engineering Management Conference*.

Fukas, P., Bozkurt, A., Lenz, N., & Thomas, O. (2023, June). Developing a maturity assessment tool to enable the management of artificial intelligence for organizations. https://doi.org/10.1007/978-3-031-34985-0_5

Fukas, P., Rebstadt, J., Remark, F., & Thomas, O. (2021). Developing an artificial intelligence maturity model for auditing.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. https://doi.org/10.48550/ARXIV.1406.2661

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, *61*(4), 5–14. https://doi.org/10.1177/0008125619864925

Hagendorff, T. (2020). The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hartikainen, M., Väänänen, K., & Olsson, T. (2023). Towards a human-centred artificial intelligence maturity model. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544549.3585752

Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the delphi survey technique. *Journal of Advanced Nursing*, *32*(4), 1008–1015. https://doi.org/10.1046/j.1365-2648.2000.t01-1-01567.x

Hein-Pensel, F., Winkler, H., Brückner, A., Wölke, M., Jabs, I., Mayan, I. J., Kirschenbaum, A., Friedrich, J., & Zinke-Wehlmann, C. (2023). Maturity assessment for industry 5.0: A review of existing maturity models. *Journal of Manufacturing Systems*, *66*, 200–210. https://doi.org/https://doi.org/10.1016/j.jmsy.2022.12.009

Helgesson, Y. Y. L., Höst, M., & Weyns, K. (2011). A review of methods for evaluation of maturity models for process improvement. *Journal of Software: Evolution and Process*, *24*(4), 436–454. https://doi.org/10.1002/smr.560

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105. https://doi.org/10.2307/25148625

Holmström, J. (2022). From ai to digital transformation: The ai readiness framework. *Business Horizons*, *65*(3), 329–339. https://doi.org/https://doi.org/10.1016/j.bushor.2021.03.006

Hsu, C.-C., & Sandford, B. (2007). The delphi technique: Making sense of consensus. *Practical Assessment, Research and Evaluation*, *12*.

Iversen, J., Nielsen, P. A., & Norbjerg, J. (1999). Situated assessment of problems in software development. *SIGMIS Database*, *30*(2), 66–81. https://doi.org/10.1145/383371.383376

Jantunen, M., Halme, E., Vakkuri, V., Kemell, K.-K., Rousi, R., Mikkonen, T., Nguyen Duc, A., & Abrahamsson, P. (2021). Building a maturity model for developing ethically aligned ai systems.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, *1*. https://doi.org/10.1038/s42256-019-0088-2

Katz, A., Balaouras, S., Sridharan, S., Megan, E., & Barton, J. (2023, October). *Inquiry spotlight: Generative ai, 2023* (tech. rep.). Forrester. https://www.forrester.com/report/inquiry-spotlight-generative-ai-2023/RES179999?ref_search=4070805_1720438403187

Keeney, S., Hasson, F., & McKenna, H. P. (2001). A critical review of the delphi technique as a research methodology for nursing. *International Journal of Nursing Studies*, *38*(2), 195–200. https://doi.org/10.1016/s0020-7489(00)00044-4

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. https://doi.org/10.48550/ARXIV.1312.6114

Krijger, J., Thuis, T., de Ruiter, M., Ligthart, E., & Broekman, I. (2022). The ai ethics maturity model: A holistic approach to advancing ethical data science in organizations. *AI and Ethics*, *3*(2), 355–367. https://doi.org/10.1007/s43681-022-00228-7

Kulkarni, U., & Freeze, R. (2004). *Development and validation of a knowledge management capability assessment model*.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy ai: From principles to practices. *ACM Computing Surveys*, *55*(9), 1–46. https://doi.org/10.1145/3555803

Lichtenthaler, U. (2020). Five maturity levels of managing ai: From isolated ignorance to integrated intelligence. *Journal of Innovation Management*, *8*(1). https://doi.org/10.24840/2183-0606_008.001_0005

Lourenço, J. M. (2021). *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. https://github.com/joaomlourenco/novathesis/raw/main/template.pdf

Madaio, M., Kapania, S., Qadri, R., Wang, D., Zaldivar, A., Denton, R., & Wilcox, L. (2024). Learning about responsible ai on-the-job: Learning pathways, orientations, and aspirations. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3630106.3658988

Martinek-Jaguszewska, K., & Rogowski, W. (2022). Development and validation of the business process automation maturity model: Results of the delphi study. *Information Systems Management*, *40*(2), 169–185. https://doi.org/10.1080/10580530.2022.2071506

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine*, *27*(4), 12–14.

Meijer, A., Lorenz, L., & Wessels, M. (2021). Algorithmization of bureaucratic organizations: Using a practice lens to study how context shapes predictive policing systems. *Public Administration Review*, *81*(5), 837–846. https://doi.org/10.1111/puar.13391

Mettler, T., & Rohner, P. (2009). Situational maturity models as instrumental artifacts for organizational design. *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology - DESRIST '09*. https://doi.org/10.1145/1555619.1555649

Ministry of the Interior and Kingdom Relations. (2024). Overheidsbrede visie generatieve al. https://open.overheid.nl/documenten/9aa7b64a-be51-4e6a-ad34-26050b8a67ef/file

Moor, J. H. (1985). What is computer ethics? *Metaphilosophy*, *16*(4), 266–275.

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as a service: A pragmatic operationalisation of ai ethics. *Minds and Machines*, *31*(2), 239–256. https://doi.org/10.1007/s11023-021-09563-w

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available ai ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, *26*(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Munn, L. (2022). The uselessness of ai ethics. *AI and Ethics*, *3*(3), 869–877. https://doi.org/10.1007/s43681-022-00209-w

Mylrea, M., & Robinson, N. (2023). Artificial intelligence (ai) trust framework and maturity model: Applying an entropy lens to improve security, privacy, and ethical ai. *Entropy*, *25*(10). https://doi.org/10.3390/e25101429

Noymanee, J., Iewwongcharoen, B., & Theeramunkong, T. (2022). Artificial intelligence maturity model for government administration and service. https://doi.org/10.1109/DGTi-CON53875.2022.9849184

OECD. (2019). Recommendation of the council on artificial intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Okoli, C., & Pawlowski, S. D. (2004). The delphi method as a research tool: An example, design considerations and applications. *Information amp; Management*, *42*(1), 15–29. https://doi.org/10.1016/j.im.2003.11.002

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson,

E., McDonald, S., . . . Moher, D. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*. https://doi.org/10.1136/bmj.n71

Paré, G., Cameron, A.-F., Poba-Nzaou, P., & Templier, M. (2013). A systematic assessment of rigor in information systems ranking-type delphi studies. *Information amp; Management*, *50*(5), 207–217. https://doi.org/10.1016/j.im.2013.03.003

Parliament, E., & of the European Union, C. (2024). 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). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

Parrish, A. H., & Sadera, W. A. (2018). Teaching competencies for student-centered, one-to-one learning environments: A delphi study. *Journal of Educational Computing Research*, *57*(8), 1910–1934. https://doi.org/10.1177/0735633118816651

Paulk, M. C., Curtis, B., Chrissis, M. B., & Weber, C. V. (1993). Capability maturity model, version 1.1. *IEEE Software*, *10*(4), 18–27. https://doi.org/10.1109/52.219617

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research [doi: 10.2753/MIS0742-1222240302]. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Pfeffer, J., & Sutton, R. I. (1999). Knowing "what" to do is not enough: Turning knowledge into action. *California Management Review*, *42*(1), 83–108. https://doi.org/10.1177/000812569904200101

Pöppelbuß, J., & Röglinger, M. (2011). What makes a useful maturity model? a framework of general design principles for maturity models and its demonstration in business process management. *European Conference on Information Systems*.

Pouliakas, K., Russo, G., Santangelo, G., et al. (2024). *Untangling labour shortages in europe: Unmet skill demand or bad jobs?* Publications Office of the European Union.

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2015). A taxonomy of evaluation methods for information systems artifacts. *Journal of Management Information Systems*, *32*(3), 229–267. https://doi.org/10.1080/07421222.2015.1099390

Proenca, D. (2016). Methods and techniques for maturity assessment. *2016 11th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–4. https://doi.org/10.1109/cisti.2016.7521483

Qiang, V., Rhim, J., & Moon, A. (2023). No such thing as one-size-fits-all in ai ethics frameworks: A comparative case study. *AI SOCIETY*. https://doi.org/10.1007/s00146-023-01653-w

Reichl, G., & Rudolf, G. (2023). Maturity models for the use of artificial intelligence in enterprises: A literature review, 486–502. https://doi.org/10.24867/IS-2023-VP1.1-5_07341

Reis, T. L., Mathias, M. A. S., & de Oliveira, O. J. (2017). Maturity models: Identifying the state-of-the-art and the scientific gaps from a bibliometric study. *Scientometrics*, *110*(2), 643–672. https://doi.org/10.1007/s11192-016-2182-0

Rowe, G., & Wright, G. (1999). The delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting*, *15*(4), 353–375. https://doi.org/10.1016/s0169-2070(99)00018-7

Sadiq, R. B., Safie, N., Abd Rahman, A. H., & Goudarzi, S. (2021). Artificial intelligence maturity model: A systematic literature review. *PeerJ Computer Science*, *7*, e661. https://doi.org/10.7717/peerj-cs.661

Salah, D., Paige, R., & Cairns, P. (2014). An evaluation template for expert review of maturity models. In *Product-focused software process improvement* (pp. 318–321). Springer International Publishing. https://doi.org/10.1007/978-3-319-13835-0_31

Saldaña, J. (2016). *The coding manual for qualitative researchers*. SAGE Publications.

Saltz, J. S., & Dewar, N. (2019). Data science ethical considerations: A systematic literature review and proposed project framework. *Ethics and Information Technology*, *21*(3), 197–208. https://doi.org/10.1007/s10676-019-09502-5

Schaschek, M., & Engel, S. (2023). Measuring trustworthiness of ai systems: A holistic maturity model.

Schmidt, M., Marrone, S., Paraschakis, D., & Singh, T. (2022). Artificial intelligence in the energy transition for solar photovoltaic small and medium-sized enterprises. In *Digital humanism* (pp. 105–123). Springer International Publishing. https://doi.org/10.1007/978-3-030-97054-3_7

Schuh, G., Anderl, R., Gausemeier, J., Ten Hompel, M., Wahlster, W., & Herbert Utz, V. (2017). *Industrie 4.0 maturity index die digitale transformation von unternehmen gestalten*. Herbert Utz Verlag GmbH.

Schumacher, A., Erol, S., & Sihn, W. (2016). A maturity model for assessing industry 4.0 readiness and maturity of manufacturing enterprises. *Procedia CIRP*, *52*, 161–166. https://doi.org/https://doi.org/10.1016/j.procir.2016.07.040

Schuster, T., & Waidelich, L. (2022). Maturity of artificial intelligence in smes: Privacy and ethics dimensions. In *Ifip advances in information and communication technology* (pp. 274–286). Springer International Publishing. https://doi.org/10.1007/978-3-031-14844-6_22

Schuster, T., Waidelich, L., & Volz, R. (2021). Maturity models for the assessment of artificial intelligence in small and medium-sized enterprises. In *Digital transformation* (pp. 22–36). Springer International Publishing. https://doi.org/10.1007/978-3-030-85893-3_2

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., et al. (2023). Practices for governing agentic ai systems. *Research Paper, OpenAI*. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf

Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Transactions on Interactive Intelligent Systems*, *10*(4), 1–31. https://doi.org/10.1145/3419764

Skinner, R., Nelson, R. R., Chin, W. W., & Land, L. (2015). The delphi method research strategy in studies of information systems. *Communications of the Association for Information Systems*, *37*. https://doi.org/10.17705/1cais.03702

Skulmoski, G. J., Hartman, F. T., & Krahn, J. (2007). The delphi method for graduate research. *Journal of Information Technology Education: Research*, *6*, 001–021. https://doi.org/10.28945/199

Sonntag, M., Mehmann, S., Mehmann, J., & Teuteberg, F. (2024). Development and evaluation of a maturity model for ai deployment capability of manufacturing companies. *Information Systems Management*, 1–31. https://doi.org/10.1080/10580530.2024.2319041

Strasser, A. (2017). Delphi method variants in information systems research: Taxonomy development and application. *Electronic Journal of Business Research Methods*, *15*.

Tarhan, A., Turetken, O., & Reijers, H. A. (2016). Business process maturity models: A systematic literature review. *Information and Software Technology*, *75*, 122–134. https://doi.org/https://doi.org/10.1016/j.infsof.2016.01.010

Teo, T. S. H., & King, W. R. (1997). Integration between business planning and information systems planning: An evolutionary-contingency perspective [doi: 10.1080/07421222.1997.11518158]. *Journal of Management Information Systems*, *14*(1), 185–214. https://doi.org/10.1080/07421222.1997.11518158

Thuan, N. H., Drechsler, A., & Antunes, P. (2019). Construction of design science research questions. *Communications of the Association for Information Systems*, 332–363. https://doi.org/10.17705/1cais.04420

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, *14*, 207–222. https://doi.org/10.1111/1467-8551.00375

Uren, V., & Edwards, J. S. (2023). Technology readiness and the organizational journey towards ai adoption: An empirical study. *International Journal of Information Management*, *68*, 102588. https://doi.org/10.1016/j.ijinfomgt.2022.102588

Vakkuri, V., Jantunen, M., Halme, E., Kemell, K.-K., Nguyen-Duc, A., Mikkonen, T., & Abrahamsson, P. (2021). Time for ai (ethics) maturity model is now. https://doi.org/10.48550/ARXIV.2101.12701

van Steenbergen, M., Bos, R., Brinkkemper, S., van de Weerd, I., & Bekkers, W. (2010). The design of focus area maturity models. *Lecture Notes in Computer Science*, 317–332. https://doi.org/10.1007/978-3-642-13335-0_22

van Wanroij, A. (2023, December). https://eenvandaag.avrotros.nl/item/meerderheid-gemeenten-gebruikt-chatgpt-en-bijna-de-helft-weet-niet-wat-hun-medewerkers-ermee-doen/

Vassilakopoulou, P., Parmiggiani, E., Shollo, A., & Grisot, M. (2022). Responsible ai: Concepts, critical perspectives and an information systems research agenda. *Scand. J. Inf. Syst.*, *34*, 3. https://api.semanticscholar.org/CorpusID:256195536

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. https://doi.org/10.48550/arxiv.1706.03762

Venkatesh, Morris, Davis, & Davis. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425. https://doi.org/10.2307/30036540

vom Brocke, J., Hevner, A., & Maedche, A. (2020). Introduction to design science research. In J. vom Brocke, A. Hevner, & A. Maedche (Eds.), *Design science research. cases* (pp. 1–13). Springer International Publishing. https://doi.org/10.1007/978-3-030-46781-4_1

Vuletić, M., Prenzel, F., & Cucuringu, M. (2024). Fin-gan: Forecasting and classifying financial time series via generative adversarial networks. *Quantitative Finance*, *24*(2), 175–199. https://doi.org/10.1080/14697688.2023.2299466

Webster, J., & Watson, R. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, *26*. https://doi.org/10.2307/4132319

Wendler, R. (2012). The maturity of maturity model research: A systematic mapping study. *Information and Software Technology*, *54*(12), 1317–1339. https://doi.org/https://doi.org/10.1016/j.infsof.2012.07.007

Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering* [10.1007/978-3-662-43839-8]. Springer. https://doi.org/10.1007/978-3-662-43839-8

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. M. (2013). Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems*, *22*(1), 45–55. https://doi.org/10.1057/ejis.2011.51

Worrell, J. L., Di Gangi, P. M., & Bush, A. A. (2013). Exploring the use of the delphi method in accounting information systems research. *International Journal of Accounting Information Systems*, *14*(3), 193–208. https://doi.org/10.1016/j.accinf.2012.03.003

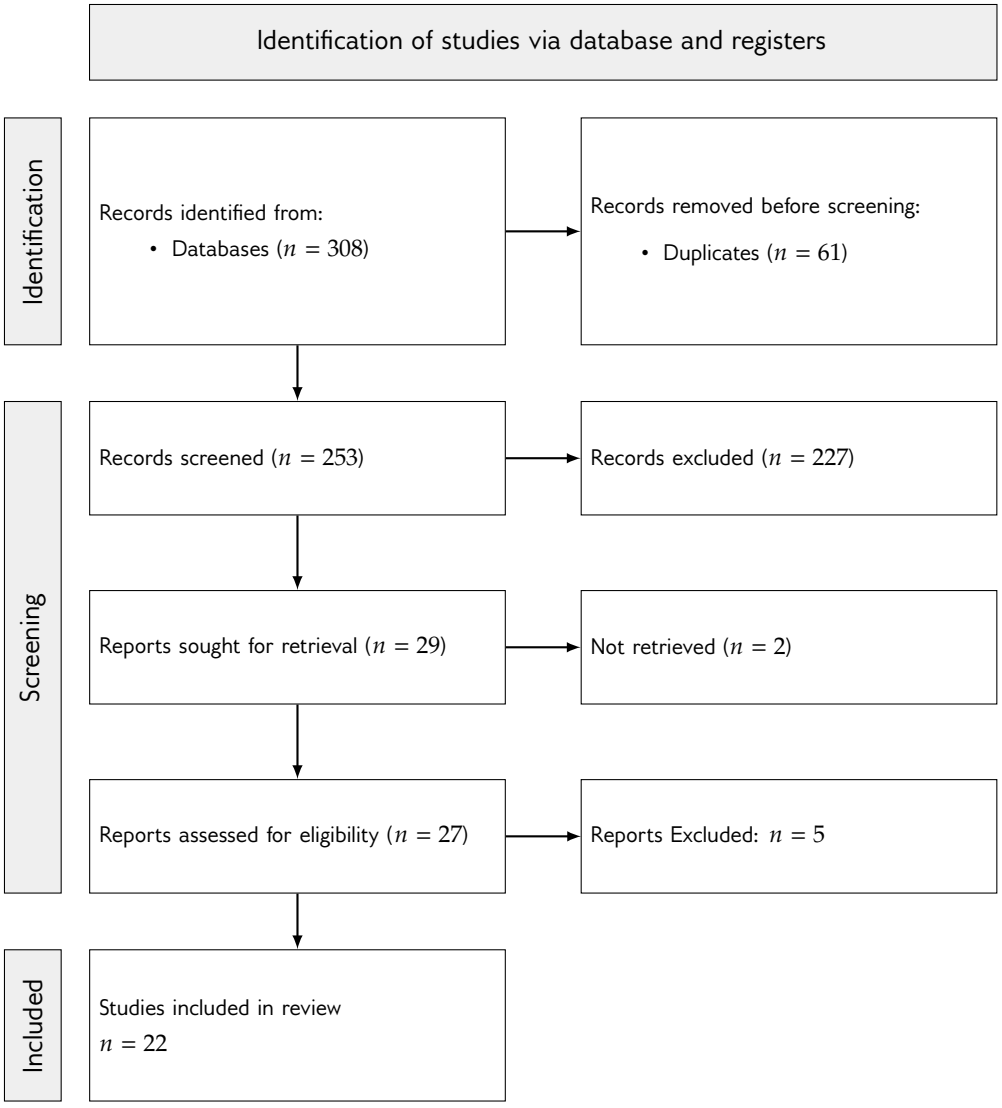
Yams, N. B., Richardson, V., Shubina, G. E., Albrecht, S., & Gillblad, D. (2020). Integrated ai and innovation management: The beginning of a beautiful friendship. *Technology Innovation Management Review*, *10*, 5–18. https://doi.org/http://doi.org/10.22215/timreview/1399

Zhobe, A., Jahankhani, H., Fong, R., Elevique, P., & Baajour, H. (2021, May). The magic quadrant: Assessing ethical maturity for artificial intelligence. https://doi.org/10.1007/978-3-030-68534-8_19

Zhou, J., & Chen, F. (2022). Ai ethics: From principles to practice. *AI amp; SOCIETY*, *38*(6), 2693–2703. https://doi.org/10.1007/s00146-022-01602-z

A

PRISMA flow chart



B

Comparison of existing maturity models

Article	Type	Model name	Model focus	Composition
Akkiraju et al. (2020)	Conference paper	Machine Learning Maturity Framework	General Structured model (CMMI)	
Alsheibani et al. (2019)	Conference paper	Artificial Intelligence	Maturity Model (AIMM)	General CMMI
Cho et al. (2023)	Article	AI Maturity Model (AI-MM)	General	Structured model (SPICE (ISO/IEC 15504))
Coates and Martin (2019)	Article	Bias governance maturity model	AI bias governance	Likert-scale questionnaires
Dotan et al. (2024)	Article	NIST AI Risk Management Framework	AI ethics	Likert-scale questionnaires
Ellefsen et al. (2019)	Conference paper	AI Maturity Model Framework	Industry 4.0	N/A
Ferreira et al. (2023)	Conference paper	RAI Maturity Model	Responsible AI	Structured model (CMMI)
Fukas et al. (2021)	Conference paper	Auditing Artificial Intelligence Maturity Model (A-AIMM)	Auditing	Hybrid
Hartikainen et al. (2023)	Conference paper	Human-Centered AI Maturity Model (HCAI-MM)	Human-Centered AI	Maturity grid
Holmström (2022)	Article	AI readiness framework	General	Maturity grid
Jantunen et al. (2021)	Conference paper	AI ethics maturity model	AI ethics	Maturity grid
Krijger et al. (2022)	Article	AI ethics maturity model	AI ethics	Maturity grid
Lichtenthaler (2020)	Article	AI management framework	AI management	N/A
Mylrea and Robinson (2023)	Article	AI Trust Framework and Maturity Model (AI-TMM)	General	Maturity grid
Noymanee et al. (2022)	Conference paper	AI Maturity Model	Government	Maturity grid
Schaschek and Engel (2023)	Conference paper	Trustworthy AI maturity model (TAI-MM)	Trustworthy AI	Maturity grid
Schmidt et al. (2022)	Book chapter	AI maturity transition framework	SMEs	Maturity grid
Schuster et al. (2021)	Conference paper	AI Maturity Model (AIMM)	SMEs	Maturity grid
Sonntag et al. (2024)	Article	SMMT maturity model	Manufacturing companies	Maturity grid
Uren and Edwards (2023)	Article	PPTD innovation journey	General	Structured model (TRL)
Yams et al. (2020)	Article	AI Innovation Maturity Index (AIMI)	General	Structured model (ISO 56002)
Zhobe et al. (2021)	Conference paper	AI ethics quadrant	General	Maturity grid

Table B.1: Structural comparison of maturity models

Article	Level	Names	Dimension	Reference	Names	Empirical?	Model type	Architecture	Validation?
Akkiraju et al. (2020).	5	Initial, repeatable, defined, managed, optimizing	9	capabilities	AI Model Goal Setting; Data Pipeline Management; Feature Preparation Pipeline; Train Pipeline Management; Test Pipeline Management; Model Quality, performance and model management; Model Error Analysis, Model Fairness and Trust; Model Transparency	No	Descriptive	Continuous	No
Alsheibani et al. (2019)	5	Initial, assessing, determined, managed, optimise	4	dimensions	AI functions; Data structure; People; Organisational	No	Descriptive	Continuous	No
Cho et al. (2023)	6	Incomplete, performed, managed, established, predictable, optimizing	13	processes	Software requirements analysis; software architecture design; data collection; data cleaning; data preprocessing; training process management; performance evaluation of AI model; Safety evaluation of AI model; Final AI model management; System safety evaluation; System safety preparedness; AI infrastructure; AI model operation management	Yes	Prescriptive	Continuous	Yes
Coates and Martin (2019)	5	N/A	11	constructs	"Design and development: business, people, user, data, algorithm, compliance; Post development: business data, testing, client feedback, compliance"	Yes	Prescriptive	Staged	Yes
Dotan et al. (2024)	5	N/A	4	pillars	Map; Govern; Measure; Manage	Yes	Prescriptive		
Ellefsen et al. (2019)	4	Novice, Ready, Proficient, and Advanced	N/A	N/A	Strategy, Organization, Data, Technology, and Operations.	No	Descriptive	Continuous	No
Ferreira et al. (2023)	4	Unaware; Exploratory; Proactive; Strategic	7	requirements	Human agency and oversight; Technical robustness and safety; privacy and data governance; Transparancy; Diversity, non-discrimination and fairness; Societal and environmental wellbeing; Accountability	Yes	Prescriptive	Continuous	No

Fukas et al. (2021)	5	Initial, assessing, determined, managed, optimized	8	dimensions	Technologies, data, people & competences, organisation & Processes, strategy & management, budget, products & services, ethics & regulations	Yes	Prescriptive	Continuous	Yes
Hartikainen et al. (2023)	3	N/A	6	criteria	Explainability; Transparency; Fairness; Accountability; collaboration and human control; Working with AI's uncertainty	No	Descriptive	Continuous	No
Holmström (2022)	5	None, Low, Moderate, High, Excellent	4	dimensions	Technologies; Activities; Boundaries; Goals	No	Prescriptive	Continuous	No
Jantunen et al. (2021)	5	Ad hoc, optimized (rest has not been defined)	9	requirements	Example requirements: understanding stakeholders, accountability, data privacy, fairness, human agency, safety & security, system oversight, transparency, well-being	No	Descriptive	Continuous	No
Krijger et al. (2022)	5	Level 1; Level 2; Level 3; Level 4; Level 5	6	dimensions	Awareness & culture, policy, governance, communication & training, development processes, tooling	No	Descriptive	Continuous	No
Lichtenthaler (2020)	5	Initial intent, independent initiative, interactive implementation, interdependent innovation, intregrated intelligence	N/A	N/A	N/A	No	Descriptive	Continuous	No
Mylrea and Robinson (2023)	4	No control; Partially implemented; Largely implemented; Fully implemented	7	pillars	Explainability; Data privacy; Technical Robustness and Safety; Transparency; Data use and design; Societal well-being; Accountability	No	Descriptive	Continuous	No
Noymanee et al. (2022)	5	Rookie level; Beginner level; Operational level; Expert level; Master level	4	aspects	Strategy; Organisation; Information; Technology	No	Descriptive	Continuous	Yes
Schaschek and Engel (2023)	5	Being aware, taking first steps, approaching strategically, operationalising, innovating	4	dimensions	Data, technology, people & culture, processes	No	Descriptive	Continuous	No
Schmidt et al. (2022)	5	Exploring, Experimenting, Formalising, Optimising, Transforming	5	dimensions	Strategy, data, technology, people, governance	No	Prescriptive	Staged	Yes

Schuster et al. (2021)	5	Novice; Explorer; User; Innovator; Pioneer	7	dimensions	Strategy; Organisation; Culture/Mindset; Technology; Data; Privacy; Ethics	No	Descriptive	Continuous	Yes
Sonntag et al. (2024)	5	Initial; Experimental; Practicing; Integrated; Transformed	5	dimensions	Culture & competencies, strategy, data, organisation, processes, technology	Yes	Prescriptive	Continuous	Yes
Uren and Edwards (2023)	3	Laying the foundations of AI: Adoption of AI; Mature AI;	4	components	People; Process; Technology; Data	No	Descriptive	Staged	No
Yams et al. (2020)	5	Foundational; Experimenting; Operational; Inquiring; Integrated	6	dimensions	Strategy; Ecosystems; Mindsets; Organisation; Data; Technology	No	Descriptive	Continuous	No
Zhobe et al. (2021)	4	The starter; The aspiring; The equipped; The leader	2	categories	Capabilities; Vision	No	Descriptive		

Table B.2: Structural comparison of maturity models continued

C

Purpose and scope of previous maturity models

Author	Purpose
Akkiraju et al. (2020)	introduce a maturity framework tailored for machine learning processes, acknowledging the distinct lifecycle of ML models compared to traditional software. Unlike the deterministic nature of software, machines learning models are probabilistic, require training and iterative improvement, and may not reach perfect accuracy. This complexity demands a new Capability Maturity Model (CMM) for effective ML model management. The maturity model is primarily descriptive and has many capabilities that vaguely describing reaching a mature ML lifecycle within enterprises.
Alsheibani et al. (2019)	were among the the first academics to propose an AI maturity model. The model combines AI functions, data structure, people, and organisational dimensions with a five-level maturity scale, drawing inspiration from established maturity models such as the Capability Maturity Model Integration (CMMI). The initial version of the AI maturity model provides a starting point for the model evaluation.
Cho et al. (2023)	introduced a new maturity model for trustworthy AI software, AI-MM, in line with the traditional ISO 15504 SPICE framework. This model comprehensively addresses common AI and quality-specific processes, focusing on incorporating new quality-specific processes related to fairness and safety. The AI-MM is designed to be an adaptable maturity model featuring indicators for standard AI base practices and fairness and safety base practices. The model encompasses 53 practices spanning common, fair, and safety domains.

Coates and Martin (2019)	introduced the Bias Governance Maturity Model, a comprehensive tool designed to effectively evaluate an organisation's ability to govern AI bias. The model offers a structured approach to assess capabilities across the phases of system creation: Design & Development and Post-Development. The first phase employs business, people, user, data, algorithm, and compliance, while the latter phase focuses on business, data, test, client feedback, and transparency. What is unique about the model of Coates and Martin (2019) is the inclusion of personnel capability maturity, recognising that AI development is a contextually specific combination of norms, technical systems, and strategic interests. Unlike most IT maturity models, which tend to emphasise only the technological aspect, this model comprises 11 constructs and 47 items, evaluated through 5-point maturity scales.
Ellefsen et al. (2019)	explored AI readiness in logistics companies, building upon a previously developed Logistics 4.0 maturity model. The authors proposed a set of AI maturity levels and conducted a survey based on these levels to ascertain the current state of AI development and maturity. They concluded that the industry is still predominantly in the early stages of adopting AI applications, categorised as the novice maturity stage. Notably, the article suggests maturity levels but does not provide dimensions or guidelines for measuring the maturity levels of companies.
Ferreira et al. (2023)	introduced a Maturity Model for Responsible AI, which encompasses a self-assessment tool and delineates the requirements, methods, and key practices essential for achieving Trustworthy AI. The self-assessment tool, consisting of 58 statements, is inspired by the Assessment List for Trustworthy AI (ALTAI) and the Ethical OS toolkit checklist. The authors followed the design process established by de Bruin et al. (2005), conducting pre-tests in two organisations and refining the model once. However, the paper does not clarify whether experts have provided feedback on the model's levels and dimensions or if any enhancements were made based on such feedback.
Fukas et al. (2023)	developed and evaluated the Auditing Artificial Intelligence Maturity Model (A-AIMM), which is designed to assess the integration and effectiveness of AI within the auditing sector, mindful of the audit process's unique demands. The model is organised around eight dimensions—Technologies, Data, People & Competences, Organisation & Processes, Strategy & Management, Budget, Products & Services, and Ethics & Regulations—and spans five maturity levels: Initial, Assessing, Determined, Managed, and Optimised. The Ethics & Regulations dimension is particularly distinctive, setting this model apart from other AI Maturity Models (Fukas et al., 2021). Following the procedural model of Becker et al. (2009), the development process culminated in a self-assessment tool. An accompanying web-based tool calculates maturity scores for each dimension, presents them in a radar chart, and generates customised action plans, thus facilitating responsible AI integration in auditing.

Hartikainen et al. (2023)	introduced a Human-Centred AI Maturity Model (HCAI-MM). The objective of Human-Centred AI is to prioritise humans over technology at the centre of AI development. The HCAI-MM serves a dual purpose: it synthesises and communicates the fundamental components of HCAI to AI developers and provides actionable guidance to address HCAI requirements, including links to existing tools. The article presents an initial outline of the maturity model, detailing the dimensions of explainability, transparency, fairness, accountability, collaboration & human control, and managing AI's uncertainty. However, the further refinement of these dimensions is not specified, and the model has yet to establish any maturity levels.
Holmström (2022)	developed an AI readiness model to assess an organisation's capability for digital transformation by implementing AI technologies. The framework focuses on four key dimensions: technologies, activities, boundaries, and goals. Organisations can evaluate their current and future potential capabilities in each dimension. Participants assess their readiness in a workshop setting, where they rate each dimension on a scale from 0 to 4. The workshops were facilitated by the author himself when applying the framework.
Jantunen et al. (2021)	propose developing an AI ethics maturity model that incorporates a scope focusing on assessing the ethical maturity of processes in AI system development. This model extends the discussions from an earlier workshop paper by Vakkuri et al. (2021). The initial model outlines five maturity levels, with only the first (ad hoc) and the last (optimised) currently defined. Examples of dimensions include understanding stakeholders, accountability, and data privacy, which still need to be finalised.
Krijger et al. (2022)	developed an AI ethics maturity model that offers a holistic framework for operationalising AI ethics within organisations. The model consists of six dimensions: Awareness & Culture, Policy, Governance, Communication & Training, Development Processes and Tooling each with five levels of maturity. It was developed based on Mutual Learning Sessions and a literature review. Although the model requires further validation in organisations, it is an initial tool applicable to public and private organisations.
Lichtenthaler (2020)	developed an AI management framework that outlines five maturity levels, offering a systematic method for assessing AI maturity in companies. The framework highlights the importance of an intelligence-based perspective on firm performance, encompassing the interplay among human intelligence, artificial intelligence, and meta-intelligence. The extent of this interaction correlates with the maturity levels as delineated in the paper.

Mylrea and Robinson (2023)	introduce the AI Trust Framework and Maturity Model (AI-TMM), which employs an entropy lens to enhance the design and governance of AI/ML systems. They suggest that a system's behaviour is most predictable when the entropy of output is maximised within structural constraints. This behaviour means that AI systems should be designed to handle high levels of disorder, allowing them to adapt to unforeseen changes. The AI-TMM integrates organisational needs for security, governance, risk, and compliance and uses Maturity Indicator Levels (MILs) to assess the implementation of controls, ranging from fully implemented (MIL Score of 3) to no control (MIL Score of 0). The model is underpinned by seven pillars of trust: Explainability, Data Privacy, Robustness and Safety, Transparency, Data Use and Design, Societal Well-Being, and Accountability, which together provides a comprehensive approach to improving trust in AI systems.
Noymanee et al. (2022)	outlined a five-level, four-aspect artificial intelligence maturity model for the public sector. The model emphasises several critical factors: strategic alignment between IT and business, organisational capabilities for big data analysis, incubation of organisational processes, consideration of employee skills, differentiation between data and information technology, industry and process specificity, and the existing IT infrastructure. This conceptual model is a theoretical foundation for future research and a guide for government organisations to implement AI effectively.
Schaschek and Engel (2023)	introduced a Trustworthy AI Maturity Model (TAI-MM) that aims to enhance existing AI Maturity Models by embedding TAI principles at all maturity levels. This initial version of the TAI-MM has delineated five maturity levels across four dimensions: data, technology, people & culture, and processes. The model presents preliminary findings on essential organisational capabilities required to develop TAI systems, although they still need to be validated.
Schmidt et al. (2022)	have significantly contributed to AI maturity transitions within the context of solar PV plant-operating SMEs. The core of his research is an AI maturity stage transition framework developed through interviews with leading PV plant operators. This framework serves as a sector-specific guide for adopting a data-driven approach in managing solar PV plant-operating SMEs. The study enhances an existing AI maturity framework, which includes the dimensions of data, technology, people, and governance, by transforming it into a prescriptive model. These dimensions are used as guiding inputs to construct the transition framework.
Schuster and Waidelich (2022)	developed an AI Maturity Model for SMEs. In two articles, they followed the first five steps of the procedural model by Becker et al. (2009) (Schuster & Waidelich, 2022; Schuster et al., 2021). The model consists of the dimensions: strategy, organisation, culture/mindset, technology, data, privacy, and ethics, and has five maturity levels. The AIMM enables a targeted and supportive self-assessment based on yes/no questions. Five easy-to-answer yes/no questions were defined logically for each of the seven dimensions to match the defined maturity levels. For each "yes," a point is added to the total, which results in a corresponding maturity level.

Sonntag et al. (2024)	introduced the SMMT Maturity Model, an AI maturity model designed to assess the current state of a manufacturing company's AI deployment capabilities. The model is organised into five progressive levels, each comprising five dimensions derived from literature. A distinctive feature of the model is that each dimension—and its associated indicators—has different weightings. These dimensions have varying significance to AI systems and the specific requirements of manufacturing companies. For example, operating AI systems with high data quality is more crucial than having a well-designed data policy. Significant indicators have been developed for each dimension, each with a corresponding question. The answers to these questions range from 1 to 5, aligning with the maturity levels. The cumulative score of the indicators determines the maturity of a manufacturing company for each dimension.
Uren and Edwards (2023)	proposed a maturity model for AI adoption that integrates Technology Readiness Levels (TRL) with a tetrahedron of sociotechnical factors. The maturity model extends the People, Processes, Technology (PPT) sociotechnical model, enhancing it with 'data'. The TRL provides a benchmark measure against which expert participants could assess situations that have occurred during projects. The model aims to advance understanding of the organisational journey towards AI adoption, exploring the relevant issues at different adoption stages.
Yams et al. (2020)	introduced the AI Innovation Maturity Index (AIMI), a framework to guide organisations towards trustworthy AI integration. AIMI is structured into five levels, spanning six dimensions: data, strategy, ecosystems, mindsets, organisation, and technologies. A cross-cutting seventh dimension, trustworthiness, is woven throughout, reflecting its interdependence with the other dimensions. However, the connections between the levels and dimensions are yet to be established. AIMI aims to systematically support the integration of AI into innovation management systems, enhancing an organisation's capacity for radical innovation.
Zhobe et al. (2021)	present an ethical maturity framework for AI that aims to guide the ethical considerations throughout the AI technology lifecycle and supporting its ethical evolution.Drawing parallels with Gartner's magic quadrant, which evaluates cloud providers based on performance, growth, and capabilities, the paper applies similar principles to assess an organisation's position in terms of AI ethics. Organisations are categorised into one of four quadrants— Niche Player, Challenger, Visionary, or Leader— based on a set of devised questions that measure their current performance and capabilities in implementing ethical AI. This categorisation helps organisations understand their learning experiences within a quadrant and how it influences their approach to AI ethics.

Table C.1: Purpose and previous maturity models

D

Interview Protocol

D.1 Introduction

1. **Welcome and Introduction**
    a) Welcome the participant and thank them for their time.
    b) Briefly introduce myself and the purpose of the study.
    c) Briefly let the participant introduce him/her and his/her experience with Responsible AI in the public sector.
    d) Explain the structure of the interview and the estimated duration.
    e) Assure confidentiality and explain how the data will be used.
2. **Consent**
    a) Obtain verbal or written consent to record the interview.
    b) Confirm the participant's consent to participate in the study.

D.2 Open-Ended Questions

1. **Understanding of Responsible AI Maturity**
    a) Can you describe what Responsible AI means to you?
    b) How would you define maturity in the context of Responsible AI?
    c) What do you think are the key characteristics of a mature Responsible AI organisation?
2. **Relevant Aspects of Responsible AI**
    a) What aspects of Responsible AI do you find most relevant or important? Why?
    b) Are there any specific areas or dimensions you believe should be prioritized in a maturity model for Responsible AI?

c) Can you provide examples of practices that you consider essential for achieving maturity in Responsible AI?

3. **Challenges and Opportunities**

a) What do you see as the main challenges in achieving maturity in Responsible AI within the public sector?

b) What opportunities do you think exist for advancing Responsible AI maturity in this context?

D.3 Discussion of Initial Maturity Model

1. **Presentation of Initial Model**

a) Present the initial maturity model developed from the literature.

b) Explain the key components and dimensions of the model.

2. **Feedback on Initial Model**

a) What are your initial thoughts on this maturity model?

b) Do you think this model captures the essential aspects of Responsible AI maturity? Why or why not?

c) Are there any components or dimensions you believe are missing or need adjustment?

d) Are there any components or dimensions you believe are redundant or should be excluded?

D.4 Conclusion

1. **Summary and Thank You**

a) Thank the participant for their valuable insights and time.

2. **Next steps**

a) Inform the participant about the next steps in the study.

E

Maturity Model Round 1

Strategy	Low	High
Vision	There is no vision and there are no goals regarding AI.	There is a well-defined, communicated, and integrated vision and goals for AI across the company that support the responsible use of AI. The vision and goals are communicated to employees on a regular basis.
AI Roadmap	There is no roadmap. The organisation the organisation lacks a structured approach for advancing AI projects from initial assessment to long-term goals.	There is a comprehensive and strategic AI roadmap that outlines each phase of AI development, from initial assessment through advanced implementation and long-term planning. The roadmap is proactively managed and continuously adapted to needs.
Policy	There is no policy on how to responsible develop and use AI.	Policy on Responsible AI is widely implemented and monitored throughout the organization. There is a continuous alignment between policies that are developed on a governmental level, organisation level, and team level. The policies clearly outline when the organisation intends to use AI and what it considers important values.
AI architecture	AI is not included in the IT architecture.	AI is integrated in the existing Enterprise Architecture on all levels of the organisation.
Investment management	The organisation lacks a clear understanding of the financial requirements for AI projects, and funds are insufficient for AI implementation.	The investment capital is constantly reviewed by appropriate executives and, if necessary, adjusted to meet requirements. There are sufficient funds available for the succesful pilots to be rolled out across the organisation.

Table E.1: Item levels for Strategy dimension after Round 1

Culture & Competences	Low	High
Training	The organisation does not offer any AI-related trainings or education programmes.	There is a fully developed training module that includes a schedule for regular training for different types of users in the organization. Employees learn how to responsibly use AI, but also learn to systematically reflect on AI.
Active management support	Management is not aware of any AI initiatives and is not promoting the use of AI.	Leaders actively promote and support AI initiatives, encouraging a culture of innovation and growth. They consistently advocate for responsible AI practices and provide resources and support to advance these goals.
Knowledge management	Employees maintain their own knowledge databases and do not share them with others.	A centralised platform for building up the collected knowledge is used by all employees and is constantly being expanded with thought leadership and information on Responsible AI. Employees can find and share information via intranet pages and are encouraged to contribute to work groups, events, and communities.
Diversity	There is no diversity in development teams and the organisation does not actively promote diverse perspectives in AI development	Development teams are diverse, including a range of perspectives and backgrounds that contribute to the fairness and inclusivity of AI systems. The organisation actively promotes and supports diversity within AI teams.
AI competences	Employees lack essential competencies related to the development, use, and improvement of AI technologies.	There is a clear overview of personal competences that employees need to possess to develop, use and improve AI technologies in an organisation. All employees have at least a basic understanding of the functionalities, benefits and risk of AI.
Ambassadors	There are no employees actively promoting or advocating for Responsible AI practices. The concept of Responsible AI is not widely discussed or supported within the organization.	There are designated ambassadors or early adopters who actively promote and advocate for Responsible AI practices throughout the organization. They play a key role in spreading awareness and driving ethical AI initiatives.
Discretion	There is no Employees are expected to follow AI system recommendations without exercising independent judgment.	Employees are encouraged to exercise judgment on the output of the AI model. For higher-risk AI applications, there is always a human-in-the-loop who uses their independent judgment to make informed decisions beyond AI suggestions.

Table E.2: Item levels for Culture & Competences dimension after Round 1

Governance & Processes	Low	High
Governance structure	AI roles and responsibilities within the organization are unclear or undefined.	The organisation has a well-defined governance structure for AI, with clear roles and responsibilities assigned from leadership to individual team members. There is a Chief (responsible) AI Officer, a Chief Data Officer, and/or a Chief Ethics offer, and the AI systems are controlled and managed by a separate business unit or person. This structure supports effective management and oversight of AI initiatives.
Stakeholder engagement	Stakeholder engagement is minimal or non-existent. There is limited involvement of employees, experts, or users in the development and ongoing discussions of AI projects.	The organisation actively engages a diverse range of stakeholders, including employees at all levels, external industry experts, potential users, and people that are affected in discussions and development of AI projects. This engagement is structured and systematic, supporting inclusive and informed decision-making.
Accountability	Responsibilities and ownership are undefined, leading to unclear lines of responsibility for AI-related decisions and outcomes.	Responsibilities are assigned and communicated effectively, ensuring that individuals and teams are held accountable for the outcomes and impacts of AI initiatives. AI initiatives are actively registered on the algorithm register and it is clear for citizens and employees who is responsible for an AI initiative.
Impact assessments	There is no assessment of the impacts of AI systems. The organisation lacks a systematic approach to understanding the usefulness, risks, and benefits of AI technologies.	The organisation systematically assesses the impacts of AI systems, including their usefulness, risks, and benefits. This assessment is used to inform decision-making and ensure that AI systems deliver positive outcomes while managing potential risks effectively.
Supplier management	The organisation lacks formal procurement guidelines for responsible AI systems. There is no insight in the ethical, technical, and legal risk associated with third-party AI solutions	The organisation has a comprehensive and transparent management process for AI procurement. Clear guidelines are in place for selecting vendors, with strict requirements for transparency, data governance, and responsible AI usage. Contracts, including data processing agreements are designed.
Compliance	The organisation does not have established mechanisms for ensuring compliance with regulations and standards related to AI.	Compliance is fully integrated into the organisation's governance framework. It goes beyond meeting legal requirements, embedding ethical considerations and continuous oversight into everyday processes. The compliance is proactive, with systems in place to anticipate regulatory changes and ethical risks.

Table E.3: Item levels for Governance & Processes dimension after Round 1

Data & Information	Low	High
Data quality	Data used for training AI models is of poor quality, with significant issues in accuracy, completeness, relevance, or bias.	Data quality is monitored against performance expectations. There are robust processes in place for regularly assessing and improving data quality.
Metadata	Metadata management is not operated. Metadata is either not captured or poorly managed, leading to challenges in tracking and understanding data.	Metadata management is in place. Metadata enables access to the right data when it is needed.
Data ecosystem	The organisation lacks a cohesive data ecosystem, resulting in fragmented data management and difficulties in integrating and accessing data across applications.	The organisation has a well-integrated data ecosystem that simplifies data management and facilitates seamless use across various applications. This system ensures efficient data integration, accessibility, and usability.
Data policy	There are no formal data policies or structures in place for processing data. Data handling, privacy, and security practices are inconsistent or undefined.	The organisation has established and enforced comprehensive data policies and structures for processing data. These policies include clear guidelines for data handling, privacy, and security, ensuring consistent and responsible data management.

Table E.4: Item levels for Data & Information dimension after Round 1

Technology & Tooling	Low	High
Tooling	There is minimal or no systematic monitoring of the functionality and quality of AI systems. Testing is infrequent and not well-documented.	The functionality and quality of AI systems are constantly monitored with tests. Organisations develop and utilise tools for assessing fairness in datasets and AI models, including bias detection tools and fairness assessment frameworks, to ensure ethical AI development .
Infrastructure	The IT infrastructure is not well-adapted to support AI systems, leading to challenges in developing, deploying, or maintaining AI solutions. There is a lack of integrated tools and platforms for AI.	The organisation has a scalable, secure, and high-performance IT infrastructure that supports complex AI workloads. The infrastructure is cloud-enabled and is agile, enabling rapid adaptation to new AI developments while maintaining robust security controls.

Table E.5: Item levels for Technology & Tooling dimension after Round 1

F

Survey description Round 2

Dear participant,

Thank you for participating in this **second round (2/3)** of the Delphi study. First off, I want to thank all of you for your responses in the interview, which have been very insightful for further developing the maturity model.

The survey includes both **closed questions**, which are used to score various dimensions/items/levels, and **open questions** for providing explanations. While your insights on design choices would be greatly appreciated, please note that responding to the open questions is optional for completing the survey. The model consists of five dimensions, presented in the following order in this questionnaire:

- Strategy
- Culture & Competences
- Governance & Processes
- Data & Information
- Technology & Tooling

This survey will take approximately **30-35 minutes** to complete, and your response will be anonymized. No names will be shared with other participants or displayed in my thesis. For any questions or other inquiries, you can always reach me at xxx@student.utwente.nl or +31 6 xxxxxxxx.

G

Aggregated feedback Round 2

Item	Feedback
Vision	• Defining a vision for AI is challenging as the technology is still developing. • The AI vision should be fully integrated into the overall organizational vision,rather than existing as a separate AI vision. • Define vision and goals as 'responsible use of AI' instead of 'use of responsible AI'. So it's not the technology that is responsible but the way we use it.
Roadmap	• A roadmap can help develop and organize AI applications, but organizations may work incrementally, experimenting with specific tasks before scaling up. • An overall roadmap for all AI applications is not necessary for successful implementation.
AI architecture	• "AI architecture" should be part of the Technology or Infrastructure dimension, not Strategy. Business architecture related to AI could be part of Strategy. • AI architecture should be included in policy. • Integration with IT infrastructure depends on the specific AI application and its data needs.
Investment management	• Investment management is a resource needed to execute strategy, not a direct indicator. • Organisations have various investment priorities. They must evaluate and prioritize investments available for AI applications, which may follow a broader roadmap or be allocated incrementally for specific experiments and scaling up successful initiatives. AI investments will compete with other organizational needs. • Many organizations currently lack a comprehensive understanding of AI's capabilities and limitations. This gap makes it challenging to define clear requirements and adjust investment strategies accordingly.

Table G.1: Aggregated results Strategy dimension Round 2

Item	Feedback
Training	• AI is an 'umbrella term'. Training needs depend on where and how AI is introduced. Not all employees require the same level of training. • Include awareness programs. • What type of training is included? only technical, or also legal and awareness? • Training should address different types of AI (predictive, prescriptive, supporting or leading etc.) and their impact on decision-making. • The EU AI Act suggests AI-informed staff, but not all roles need AI training.
Active management support	• Technologists should not be siloed; leadership must integrate them with the rest of the organization. • Low awareness at the management level often results in insufficient promotion of AI competencies. • If an AI system recommends 'do A', it takes courage to 'do B'. Without strong management support, employees are unlikely to challenge AI recommendations. • Not promoting AI should not be a disqualifier. Rephrase to:not promoting the responsible use of AI.
Knowledge management	• Not all knowledge needs to be shared with all employees.
Competences	• Competences for the responsible use of AI in all phases, including auditability, are missing in the model.
Diversity	• A diverse team is critical as AI impacts people's lives differently. Important aspects may be missed without diversity. • Unclear what is meant by diversity. It is a difficult and nuanced topic. • Diversity is not always essential for every AI application; it depends on the task. • Diversity enhances innovation but is not directly linked to AI.
Discretion	• Employees should not follow AI blindly; they need the ability to accept or reject recommendations. • Discretion is important. Then again, if the staff does not follow AI recommendations, the AI is not performing well. • Discretion should be included in policy.
Ambassadors	• Ambassadors are important but not essential. • Combine ambassadors with leadership promotion under a new item called 'active support'. • Ambassadors should be broad within the organization, not just at the director level. • Ambassadors cost money and will only be allowed if they bring benefits or their actions are expected.

Table G.2: Aggregated results Culture & Competences dimension Round 2

Item	Feedback
Governance structure	• The governance structure should fit within the overall organizational governance framework, ensuring an organization is "in-control" over their processes. AI is another process they need to be in control of. • Reaching the transformative level (4) should not be limited to having specific roles like Chief AI Officer; use "for example" to suggest roles.
Stakeholder engagement	• Be cautious in defining stakeholders as this may evolve. • Clarify the distinction between governance structure and stakeholder engagement.
Accountability	• Post-implementation accountability is key as not all scenarios can be considered during training and testing. • Clarify the definition of accountability, possibly integrating it with roles and responsibilities (i.e.governance structure). • The process and system outcomes can be reviewed by an internal or external party to ensure oversight.
Compliance	• Compliance should include adherence to internal policies for responsible AI. • Include supervision (toezicht) as part of compliance.
Impact assessment	• Include risk mitigation actions • Impact assessment should be a standard project step. Soshould you not assess all project processes?
Supplier management	• Prefer in-house technology development for government use cases. At a minimum internal employees need to understand how the system works. • What part is governed? Supplier management is an operational activity. • Question if responsible AI usage is a supplier issue.

Table G.3: Aggregated results Governance & Processes dimension Round 2

Item	Feedback
Data quality	• Data quality is critical. Employees must assess their data and not rely on complex architectures to solve data issues. • The relevance of data quality depends on the type of AI system. This assumes data-driven AI. • Data quality is a broad concept and essential for AI applications to be effective. High-quality data defines AI outcomes. The other data and information aspects follow this crucial step. • The definition of high data quality should include specific quality requirements. • Data quality is broader than just AI usage.
Metadata	• Metadata is part of data quality; both need to be of high quality. Metadata can also be used in some cases to train AI models.
Data ecosystem	• The description of the data ecosystem is unclear, particularly its relation to data policy. • Question the necessity of having a data ecosystem for this dimension. • The phrase "facilitates seamless use across various applications" may not be appropriate or desirable in every situation.

Table G.4: Aggregated results Data & Information dimension Round 2

Item	Feedback
Tooling	• Not clear what is meant by tests • Tooling should assess not only functionality and quality but also fairness and non-discrimination. Add this explicitly to the description. • Include iterative re-assessment to determine if the AI should continue in its current form, be replaced, or revert to non-AI processes. • Incorporate tooling for risk management and security management.
Infrastructure	• Definition is not fit for every situation. • Safeguards against risks like unauthorized access and adversarial attacks are probably already in place within an organization and need to extend to AI applications as well.

Table G.5: Aggregated results Technology & Tooling dimension Round 2

H

Votes for each item Round 2

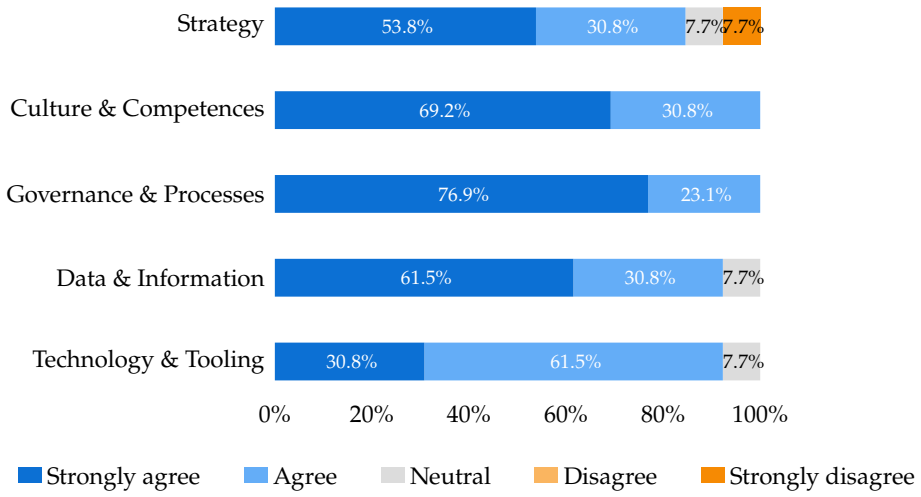


Figure H.1: Aggregated results dimensions round 2

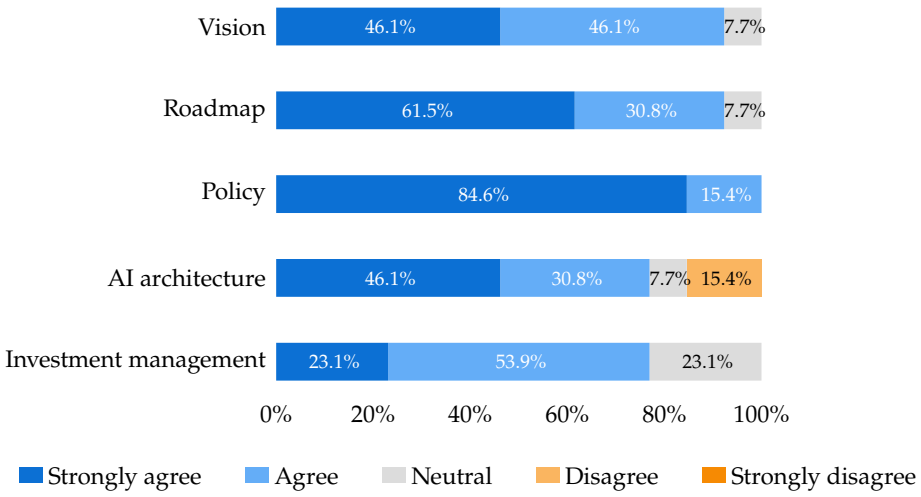


Figure H.2: Aggregated results strategy dimension round 2

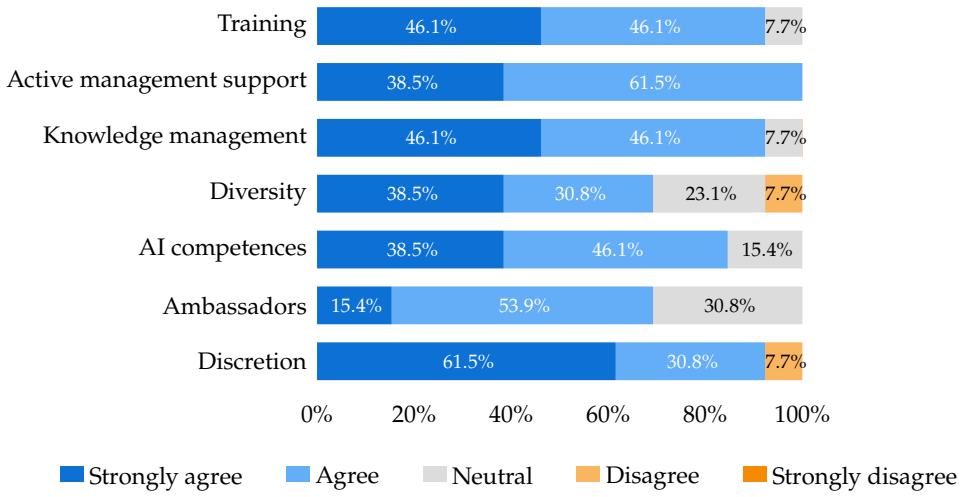


Figure H.3: Aggregated results strategy dimension round 2

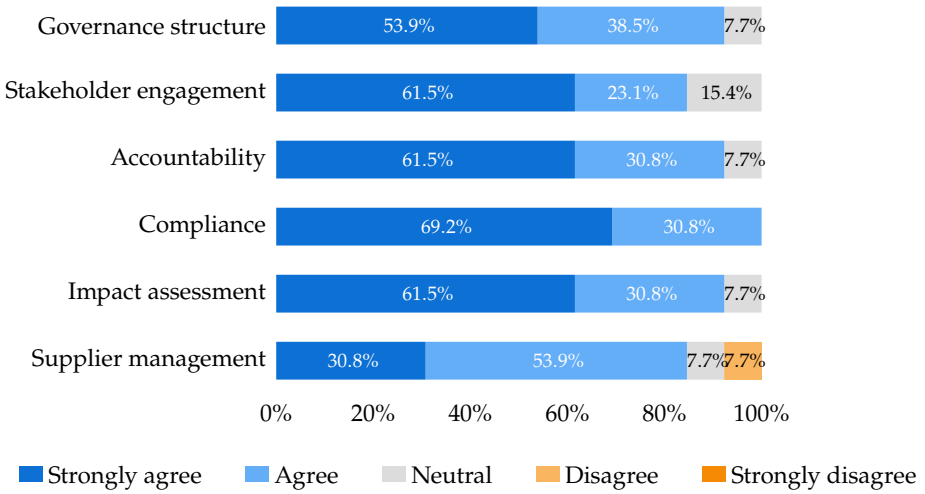


Figure H.4: Aggregated results Governance & Processes dimension round 2

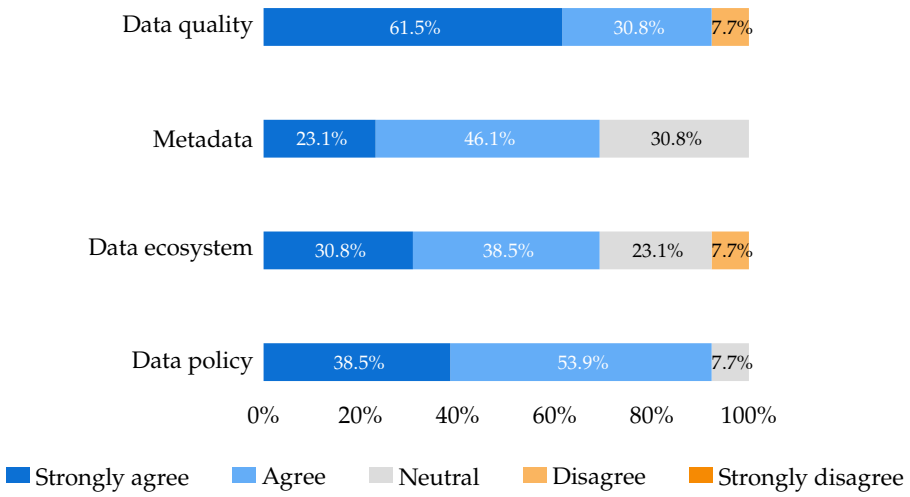


Figure H.5: Aggregated results Governance & Processes dimension round 2

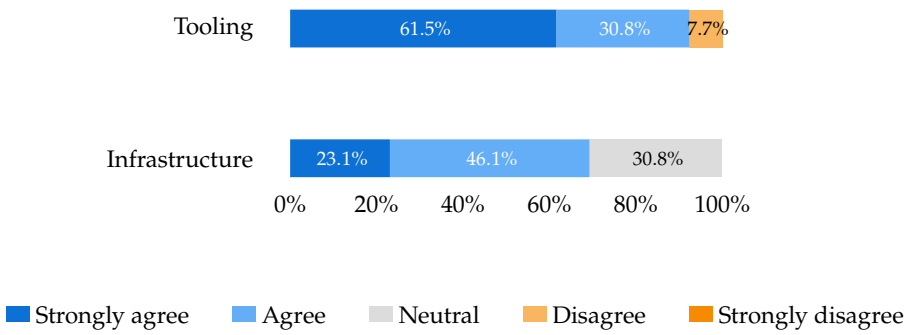


Figure H.6: Aggregated results Governance & Processes dimension round 2

I

Aggregated feedback Round 3

Item	Feedback
General	• It is also important to add something about the need to have an inventory/register of al the AI, and that there is a risk-based approach to it.
Vision	• AI supports organisational goals. This could be included in the Vision & Goals item, but that's not clear at the moment. • The following should be added to Transformed: they are also executing a formal process to maintain vision and goals. • It is not only important to communicate it to the employees, but also that the employees really adopt/live it . • There should be a process in place to change the vision when the environment changes (update it in an PDCA cycle).
Roadmap	• A strategy should have some stability long term. A roadmap is less stable and will continue to change as the AI market is not mature. • On Transformed, potentially add something about a well-defined timeline, with expected dates. • Is the roadmap not a focus on the AI projects which are required for the organisation, and how Responsible AI plays a role therein? An is the well-defined process for Responsible AI not part of the AI architecture?
AI architecture	• Surprise to see AI architecture left out with 76.9% of vote either voting 'Agree' or 'Strongly Agree'. Perhaps it is difficult where to place it, but architecting is a vital process to ensure well-defined processes, practices, and dependencies when it comes to Responsible AI development and usage.
Investment management	• A vision document could include the size of the investment over the years. However, a detailed financial plan can only be made in a next phase when the plans are more detailed. • Transformed reads as all funds are available, but that it is not how it works in practice. If the investment capital is not available, the vision & goals and roadmap will be adjusted. • Perhaps investment management should be placed under Technology & Tooling or Governance & Processes? • Combine supplier management with investment management in one item.

Table I.1: Aggregated results Strategy dimension Round 2

Item	Feedback
General	• The Culture & Competences are part of the overall organisational culture and competences. All employees should be constantly aware of how AI can help the organisation pursue its goals.
Active support	• Call active support AI Adoption and describe it as the perspect to AI, of the management/employees
Competences	• On transformed, add something about the value of competencies, that they will align with the future growth role within the organisation.
Diversity	• Diversity did not make it to the next round, but I still think it is an important aspect
Discretion	• Discretion should not be a factor from what I understand of it.
Experimentation	• Experimentation is expected to be part of 'Technology' if it is about doing proof of concepts or pilots. • Experimentation should be a technology-aspect.

Table I.2: Aggregated results Culture & Competences dimension Round 2

Item	Feedback
General	• I miss something about risk-management / risk-based processes.
Governance structure	• New description transformed: Governance structure for AI, with clear roles and responsibilities assigned from leadership to individual team members. AI systems are controlled and managed effectively with a clear oversight of AI initiatives.
Stakeholder engagement	• AI, especially Generative AI, comes so close to the primary process of an organisation, that I would suggest adding a kind of advisory group to the dedicated unit. This could be part of the stakeholder engagement. We need to avoid the situation with regular IT, where the unit is more concerned with systems than with the users and how the systems support them.
Impact assessment	• Impact assessment of what? Most impact assessments are compliance related.
Supplier management	• For public organisations, I am more supportive of doing as much AI/Machine Learning in-house as possible, especially considering the sensitivity of public sector data.

Table I.3: Aggregated results Governance & Processes dimension Round 2

Item	Feedback
General	• It is important to ad something about the need of bias scanning and coumentation of datasets, for reasons of transparancy.
Data quality	• Consider discussing the post-launch reviews to ensure data quality issues such as data drift do not arise post-production. • By stating it this way AI-initiative become responsible to fix quality issues in source system. This is a dangerous target to give somebody. The highest maturity level is achieved when: Data quality is monitored against performance expectations. There are robust processes in place for regularly assessing the data quality. Data quality metrics are regularly reviewed and acted upon. Ethical and legal consequences are deeply embedded in data quality management processes.
Metadata	• Not clear what metadata achieves in AI. The current thinking in AI is that, in some way, digitization, should be done in a vector database so that AI can reach the data effectively. This fits the Data & Information section. • Not sure if there needs to be a distinction between data management and metadatamanagement.
Data policy	• Part of policy in the strategy dimension. Policy is about the AI model and data. It could be confusing to separate those. • On Transformed, you could add that policy is constantly updated in a plan-do-check-act cycle.

Table I.4: Aggregated results Data & Information dimension Round 2

Item	Feedback
General	• Important to add something about documentation and post-market monitoring, since those are high-risk measures from the AI Act. Those two can be added to the Technology dimension.
Tooling	• Tooling is very fairness-focused, but can also be used for (i) providing transparency to all stakeholders including end-users; (ii) assessing the generalisability and/or application limits of the AI system and its outcomes; (iii) testing the security of AI systems. • I expect organisations to start with current generative AI platforms, open sourced or not. The next step will probably be to integrate the generative tools with existing IT tools. This may be too specific for the maturation model, but I believe there will be some kind of hybrid tooling. And that needs to be supported and regulated in specific ways.

Table I.5: Aggregated results Technology & Tooling dimension Round 2

Item	Feedback
General	• The current descriptions are not value-free. Some of the descriptions are negatively laden. The descriptions should be in line with the phase an organisation is in terms of AI.

Table I.6: Aggregated results Technology & Tooling dimension Round 2

J

Votes Round 3

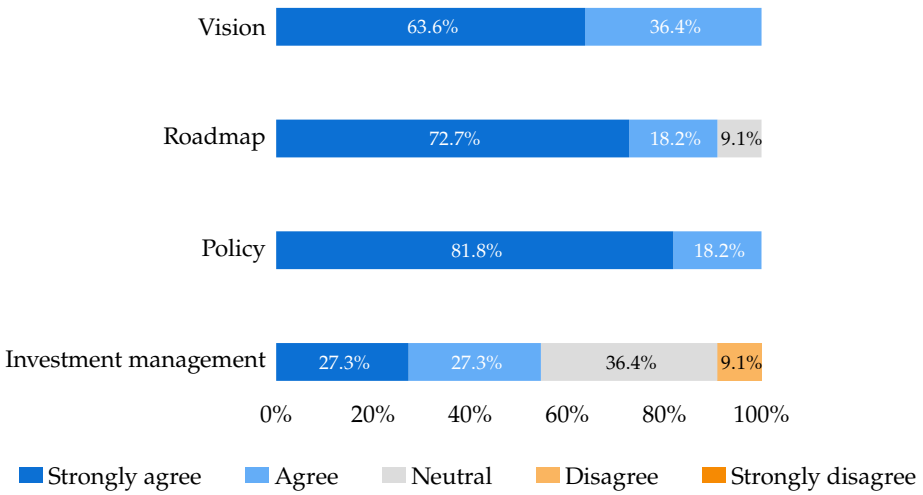


Figure J.1: Aggregated results strategy dimension round 2

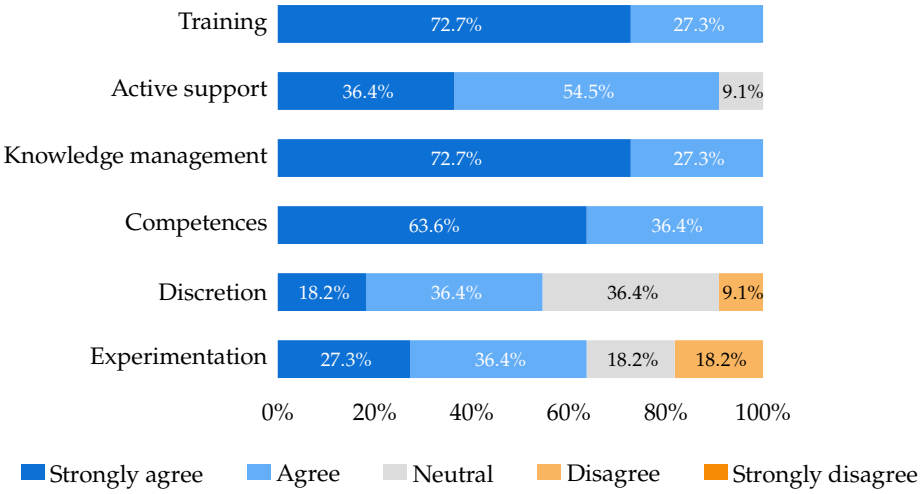


Figure J.2: Aggregated results strategy dimension round 2

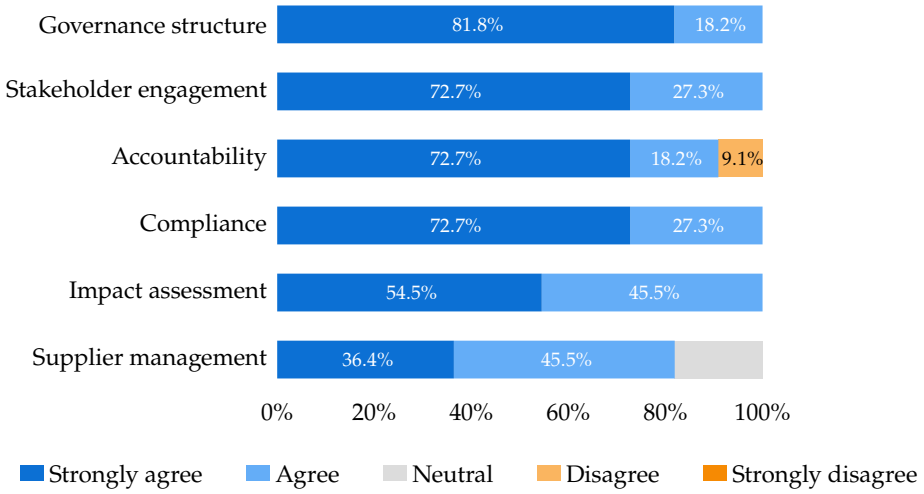


Figure J.3: Aggregated results Governance & Processes dimension round 2

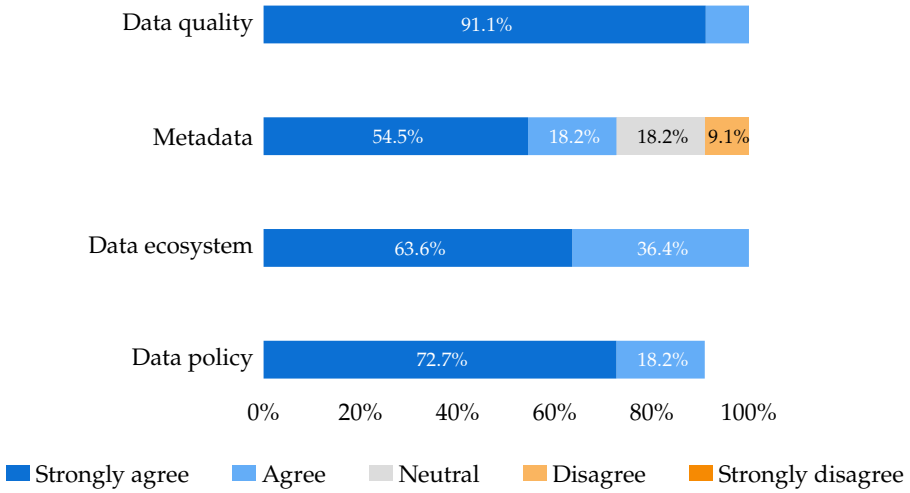


Figure J.4: Aggregated results Governance & Processes dimension round 2

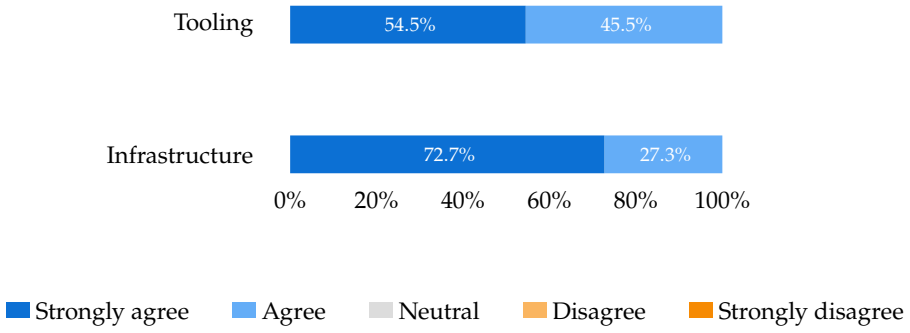


Figure J.5: Aggregated results Governance & Processes dimension round 2

K

Maturity Model Round 3

Strategy	Initial	Experimental	Practicing	Integrated	Transformed
Vision	There is no vision or goal for Responsible AI. The organisation continues to operate as usual without specific AI-related objectives.	Initial awareness of the need for a Responsible AI vision and goals, but they are not yet formalized. The distinction between strategy and tactical is not clear.	AI vision and goals are defined but not yet fully communicated or integrated across the organisation. There is a shared understanding of Responsible AI, and initial steps are taken to align AI initiatives with organisational goals.	Responsible AI vision and goals are well-defined and communicated, with alignment to broader organisational objectives. There is a process in place to update the vision and goals regularly, considering changes in the environment (PDCA cycle).	The organisation has a well-defined, communicated, and integrated vision and goals for the responsible development and/or use of AI, spanning from departmental to organisational levels. These vision and goals are regularly communicated to employees and are actively maintained through a formal process.
Roadmap	There is no roadmap. The organisation is in the early stages of exploring AI and its potential applications.	There is a growing recognition that AI can facilitate change. Initial discussions and plans are being made, but a structured roadmap is not yet in place.	A preliminary AI roadmap exists, outlining key initiatives and milestones. The roadmap begins to address identified challenges and suggests how the transition to AI should be made responsibly. It is recognised that the roadmap will change as the AI market matures.	A detailed AI roadmap is in place, with a focus on responsible AI development and/or use. This roadmap is integrated with business planning and is regularly updated. It includes a well-defined timeline with expected dates, acknowledging the dynamic nature of the AI developments.	There is a comprehensive and strategic Responsible AI roadmap that outlines each phase of Responsible AI development, from initial assessment through advanced implementation and long-term planning. The roadmap is proactively managed and continuously adapted to needs. It includes a well-defined timeline with expected dates.
Policy	There are no established policies or guidelines for the development, deployment, and use of AI technologies. AI initiatives are ad-hoc and lack oversight.	Initial policies are being drafted, focusing on basic guidelines and standards. These policies are not yet implemented or enforced. There is minimal awareness and understanding of AI policies among employees.	Policies are defined and partially implemented. Some enforcement mechanisms are in place. Practical tools and guidelines, such as quick guides, are being developed to aid in the implementation of AI policies. There is an initial alignment with organisational values and norms.	Comprehensive policies are implemented and enforced with regular reviews and updates, aligning with organisational and governmental policies. Practical tools and guidelines are widely used, and AI policies are integrated into the organisation's governance structures.	Policies on Responsible AI are widely implemented and continuously monitored. There is continuous alignment between policies developed at governmental, organisational, and team levels. The policies clearly outline when and how the organisation intends to use AI, emphasising transparency, accountability, and ethical considerations. Practical tools and guidelines, including architectural principles, are fully integrated into daily operations.

Table K.1: Item levels for Strategy dimension after Round 3

Culture & Competences	Initial	Experimental	Practicing	Integrated	Transformed
Training and awareness	The organisation does not offer any AI-related training or education programmes. Employees lack awareness of AI and its implications.	Initial awareness of the need for responsible AI training exists, but no formal programmes are in place. Training initiatives are mostly ad-hoc and occur within small teams involved in AI development or use.	Basic responsible AI training programmes are available across the organisation, focusing on foundational skills to raise employee awareness. These programmes are not regularly updated and do not cover advanced topics.	Comprehensive Responsible AI training programmes are established, regularly updated, and tailored to various roles. There is a strong emphasis on continuous education, including up-to-date knowledge of regulations and best practices. Different types of AI and their impact on decision-making are addressed. Regular awareness campaigns are conducted to keep employees informed and engaged.	A fully developed training module includes regular, role-specific training. Employees learn to develop and/or use AI responsibly and reflect on its impact. The future-proof programmes keep the organisation ahead of AI developments, covering AI risks, proper use, and legal requirements. Training aligns with the organisation's strategic goals to ensure responsible AI development and/or use.
AI adoption and support	Management is unaware of any AI initiatives and does not promote the responsible development and/or use of AI. There are no employees actively advocating for Responsible AI practices.	Management is aware of AI initiatives but offers minimal support. Some employees are aware of Responsible AI, but there are no formal ambassadors to champion these practices.	Management supports AI initiatives and begins to promote them. Initial steps are taken to identify and support Responsible AI ambassadors.	Management actively promotes AI initiatives and provides the necessary resources. Responsible AI ambassadors are identified and supported within the organisation.	Management champions AI initiatives and responsible AI practices, fostering a culture of innovation and growth. Ambassadors advocate for Responsible AI throughout the organisation, spreading knowledge and best practices.
Knowledge management	Employees maintain their own knowledge databases and do not share them with others.	Initial, informal efforts to share knowledge are in place. Some systems are being developed to facilitate access to information about Responsible AI, but they are not yet fully implemented.	Knowledge sharing practices are established and partially integrated. A centralized knowledge repository is created, allowing employees to find and share information about AI solutions, their purposes, and their authors. Employees are encouraged to contribute to work groups and events.	Knowledge sharing is actively encouraged and supported with formal processes and tools. A centralized platform is continuously expanded, and documentation about AI systems ensures continuity and responsible development and/or use. Employees regularly share knowledge internally and externally, participating in cross-organisational exchanges.	Knowledge sharing is proactively managed. The organisation actively seeks information from external sources to stay updated with the latest developments and best practices, using this information to develop thought leadership and update Responsible AI practices.

Competencies	The organisation has no understanding of the important competencies required, and employees lack essential skills related to the development and/or use of Responsible AI.	The organisation has an initial idea of relevant competencies that employees need to possess, but it is still generic, not pursued, and not consistent across the organisation.	Employees have general AI literacy, including a basic understanding of Responsible AI concepts and technologies. Competencies are starting to be tailored to specific roles.	Responsible AI competencies are clearly defined and seamlessly integrated into personnel policies. These competencies are related to understanding biases, possessing ethical expertise, legal expertise (i.e. auditability), and/or developing algorithms.	Responsible AI competencies are actively monitored and updated, according to future changes. Competencies are tailored to each employee and for each AI system, as different systems could require different competences.
Discretion	There is an over-reliance on AI, and human judgment and questioning about AI systems is minimally present.	Employees may occasionally question AI outputs, but there is no structured support for this.	Employees are encouraged to exercise discretion, but there are no formal guidelines. They are empowered to ask questions about the systems being used and to monitor outcomes for anomalies.	Formal guidelines support the exercise of discretion in AI-related decisions. A culture is fostered where employees are encouraged to exercise judgment and monitor AI outputs.	Employees are encouraged to exercise judgment on the output of the AI systems. For higher-risk AI applications, there is always a human-in-the-loop who uses their independent judgment to make informed decisions beyond AI suggestions. Employees are empowered to question and validate AI outputs, ensuring accuracy and reliability.

Table K.2: Item levels for Culture & Competences dimension after Round 3

Governance & Processes	Initial	Experimental	Practicing	Integrated	Transformed
Governance structure	AI roles and responsibilities within the organisation are unclear or undefined. There is no formal governance structure in place to oversee AI initiatives.	Some AI roles and responsibilities are defined, but they are not consistently applied or communicated. Initial steps are taken to establish a governance structure, but it lacks integration and comprehensive oversight.	AI roles and responsibilities are defined and there is some oversight. Ethical committees or boards may be established, but their roles are not fully integrated into the governance structure.	AI roles and responsibilities are well-defined and integrated within the organisation. Ethical oversight is established, and there is a multidisciplinary approach to governance. Transparent processes are in place to document and communicate AI-related decisions. The governance structure is regularly reviewed and updated.	The organisation has a well-defined and flexible governance structure for AI, with clear roles and responsibilities assigned from leadership to individual team members. AI systems are controlled and managed effectively, with a clear oversight of AI initiatives.
Stakeholder engagement	There is little to no engagement with stakeholders. Stakeholders are not involved in AI development or decision-making processes. Communication is minimal and unstructured.	Some efforts are made to engage stakeholders, but these are sporadic and not systematically integrated into AI projects. Stakeholder input is occasionally sought, but there is no formal process for incorporating their feedback.	There are defined processes for engaging stakeholders, including regular consultations and feedback mechanisms. Stakeholders from various groups, including end-users and domain experts, are involved in discussions about AI development and implementation.	Stakeholder engagement is an integral part of AI governance. There are formal structures and committees, such as ethical panels and user groups, that regularly contribute to AI projects. Stakeholder feedback is systematically incorporated into decision-making processes.	Stakeholder engagement is deeply embedded in the organisation's culture and governance. Committees, such as ethical panels and user groups, work closely with the dedicated AI unit to maintain a user-centric approach. There is continuous and proactive engagement with a wide range of stakeholders, including the public, to ensure transparency, accountability, and inclusivity. Feedback loops are well-established, and stakeholder input significantly shapes AI policies and practices.
Accountability	Responsibilities and ownership are undefined, leading to unclear lines of responsibility for AI-related decisions and outcomes. No clear process for addressing issues with AI. People affected by AI decisions have no direct way to contact the organisation.	Initial steps are taken to establish accountability, but mechanisms are weak and lack enforcement. Transparency and explainability of AI decisions are minimal. People affected by AI decisions can contact the organisation, but the process is unclear and not well-communicated.	Mechanisms for accountability are in place, but they are not fully integrated into the organisational structure. Transparency and explainability of AI decisions are partially addressed, with some processes for documenting and reviewing AI outcomes. There are some processes for people affected by AI decisions to contact the organisation, but these are not always efficient or well-publicized.	Accountability mechanisms are robust, with regular reviews and updates. Transparency, explainability, and traceability of AI decisions are prioritised, and there are established processes for addressing incidents and ethical concerns. Clear processes are established for people affected by AI decisions to contact the organization, and these are well-communicated and efficient.	Responsibilities are assigned and communicated effectively, ensuring that individuals and teams are held accountable for the outcomes and impacts of AI initiatives. AI initiatives are actively registered on the algorithm register and it is clear for citizens and employees who is responsible for an AI initiative. People affected by AI decisions have multiple clear and efficient ways to contact the organization. Transparency, explainability, and traceability of AI decisions are deeply embedded in the organisational culture, with comprehensive processes for addressing and learning from incidents.

Compliance	The organisation does not have established mechanisms for ensuring compliance with regulations, standards, or internal policies related to AI.	Some awareness of legal, ethical, and internal policy requirements exists, but compliance processes are ad-hoc and not systematically implemented. Initial steps are taken to understand and address compliance, but there is no formal structure or consistent enforcement.	Compliance processes are defined and partially implemented. There is a basic understanding of relevant laws, regulations, and internal policies, and some mechanisms are in place to ensure adherence. However, these processes are not fully integrated into the organisational structure, and enforcement is inconsistent.	Compliance processes are well-defined and integrated into the organisational structure. There is a comprehensive understanding of relevant laws, regulations, and internal policies, and robust mechanisms are in place to ensure adherence. Regular audits and reviews are conducted to maintain compliance, and there is a proactive approach to addressing potential compliance issues.	Compliance is fully integrated into the organisation's governance framework. It goes beyond meeting legal requirements, embedding ethical considerations and continuous oversight into everyday processes. The compliance is proactive, with systems in place to anticipate regulatory changes and ethical risks.
Impact assessment	There is no assessment of the impacts of AI systems. The organisation lacks a systematic approach to understanding the usefulness, risks, and benefits of AI systems.	Initial steps are taken to assess the impacts of AI systems, but the process is not yet formalised or comprehensive. Some awareness of ethical and privacy concerns exists, but there is no consistent evaluation.	Defined processes for assessing the impact of AI systems are in place and partially implemented. There is a basic understanding of the potential effects on stakeholders and society, and some mechanisms are in place to evaluate these impacts. Ethical considerations are partially integrated, but assessments are inconsistent.	Well-defined and integrated processes for assessing the impact of AI systems. Comprehensive understanding of the potential effects on stakeholders, society, and ethical standards. Regular assessments are conducted, and there is a proactive approach to identifying and mitigating negative impacts.	The organisation systematically assesses the impacts of AI systems, including their usefulness, risks, and benefits. This assessment is used to inform decision-making and ensure that AI systems deliver positive outcomes while managing potential risks effectively. There is a continuous improvement loop where assessments are regularly updated and refined to adapt to new challenges and opportunities.
Supplier management	The organisation lacks formal procurement guidelines for responsible AI systems. There is no insight in the ethical, technical, and legal risk associated with third-party AI solutions.	Basic supplier management processes are in place. Some criteria for supplier selection and evaluation are defined, but they are not consistently applied or enforced.	Supplier management processes are defined and partially integrated into the organisational structure. There is a basic understanding of the ethical, technical, and legal risks associated with third-party AI solutions.	Formal supplier management processes with clear criteria for selection, evaluation, and monitoring are in place. Regular reviews are conducted to ensure suppliers meet the organisation's standards. Ethical and technical alignment with suppliers is prioritised.	The organisation has a comprehensive and transparent process for AI procurement. Clear guidelines are in place for selecting vendors, with strict requirements for transparency, data governance, and responsible AI development and/or usage. Contracts, including data processing agreements are well-defined, and there is continuous monitoring and evaluation of supplier performance to ensure alignment with organisational values and ethical standards.

Table K.3: Item levels for Governance & Processes dimension after Round 3

Data & Information	Initial	Experimental	Practicing	Integrated	Transformed
Data quality	Data used for training AI systems is of poor quality, with significant issues in accuracy, completeness, relevance, or bias. There is no systematic approach to monitoring or improving data quality.	Initial efforts to monitor and manage data quality are underway, but they are not yet formalised. Some awareness of data quality issues exists, but there is no consistent approach to addressing them.	Data quality management processes are defined and partially implemented. There is a basic understanding of data quality parameters, and some mechanisms are in place to monitor and improve data quality. However, these processes are not fully integrated into the organisational structure, and enforcement is inconsistent.	Data quality is systematically monitored and managed. There are automated processes in place to ensure data accuracy, completeness, and consistency. Regular reviews and updates are conducted to maintain high data quality standards and to prevent data drift. Ethical and legal considerations are integrated into data quality management.	Data quality is monitored against performance expectations. There are robust processes in place for regularly assessing and improving data quality. Data quality metrics are regularly reviewed and acted upon. Ethical and legal considerations are deeply embedded in data quality management processes.
Data ecosystem	The organisation lacks a cohesive data ecosystem, resulting in fragmented data management and difficulties in integrating and accessing data across applications.	Initial steps are taken to develop a data ecosystem, but it is not yet fully integrated or coordinated. Data is still managed in silos to a large extent.	The data ecosystem is partially developed and integrated, but silos still exist. There are processes and tools in place to manage and use data, but they are not fully coordinated or optimized.	The data ecosystem is well-developed and integrated. There are robust processes and tools in place to manage and use data. Data from various sources is integrated and standardised, ensuring consistency and reliability.	The organisation has a well-integrated data ecosystem that simplifies data management across various applications. This system ensures a secure and efficient data integration, accessibility, and usability.
Data policy	There are no formal data policies or structures in place for processing data. Data handling, privacy, and security practices are inconsistent or undefined.	Initial steps are taken to develop a data policy, but it is not yet fully implemented or enforced.	A data policy is in place and partially implemented. There are rules and guidelines for managing and using data, but they are not fully enforced or integrated into the organisation's practices.	The data policy is well-developed and implemented. There are clear rules and guidelines for managing and using data, and they are regularly reviewed and updated.	The organisation has established and enforced comprehensive data policies and structures for processing data. These policies include clear guidelines for data handling, privacy, security, ownership, and sharing, ensuring consistent and responsible data management.

Table K.4: Item levels for Data & Information dimension after Round 1

Technology & Tooling	Initial	Experimental	Practicing	Integrated	Transformed
Tooling	There is minimal or no systematic monitoring of the functionality and quality of AI systems. There is no access to Responsible AI-specific tools.	Initial steps are taken to acquire and/or implement tools, but there are still significant gaps. The tools available are not fully integrated or optimised. There are some Responsible AI-specific tools, but they are limited.	The organisation has a basic set of tools to support the monitoring and assessment of AI systems. Responsible AI-specific tools are partially integrated and used, but there are still limitations in their capabilities and performance.	The organisation has a comprehensive set of Responsible AI-specific tools to support the monitoring and assessment of AI systems throughout the AI lifecycle. These tools are well-integrated and optimised, and they support the effective monitoring and assessment of AI systems.	The functionality and quality of AI systems are constantly monitored and assessed with tools and updated when necessary throughout the entire AI lifecycle. Organisations develop and/or utilise tools for assessing fairness in datasets and AI systems, providing transparency to stakeholders, assessing the generalisability and application limits of AI systems, and testing the security of AI systems.
Infrastructure	The IT infrastructure is not well-adapted to support AI systems, leading to challenges in developing, deploying, or maintaining AI solutions. There is a lack of integrated tools and platforms for AI.	Initial steps are taken to upgrade the IT infrastructure, but it is not yet fully capable of supporting AI systems.	The IT infrastructure can support basic AI systems, but there are still limitations in scalability and performance.	The IT infrastructure is fully capable of supporting advanced AI systems. It includes robust data storage, processing power, and network capabilities.	The IT infrastructure is optimized for AI systems, with advanced capabilities for data storage, processing power, and network performance. There are robust security measures in place to protect AI systems and the data they process.
Experimentation	There is minimal support for experimentation with AI technologies. Innovation is not encouraged, and there are no safe environments for testing AI technologies.	Initial steps are being taken to support experimentation with AI, but it is not yet fully implemented or encouraged. Safe environments for testing are limited.	There is a growing emphasis on responsible innovation through experiments. Employees are provided with an environment for testing.	There is dedicated support for experimentation with AI. Trust is built through extensive testing and experience. Choices are documented to ensure transparency.	The organisation fully supports and encourages experimentation with AI technologies. Innovation is a core part of the organisational culture. The duration of experimentation for AI systems is adjusted based on the risk level to ensure a responsible roll-out.

Table K.5: Item levels for Technology & Tooling dimension after Round 1