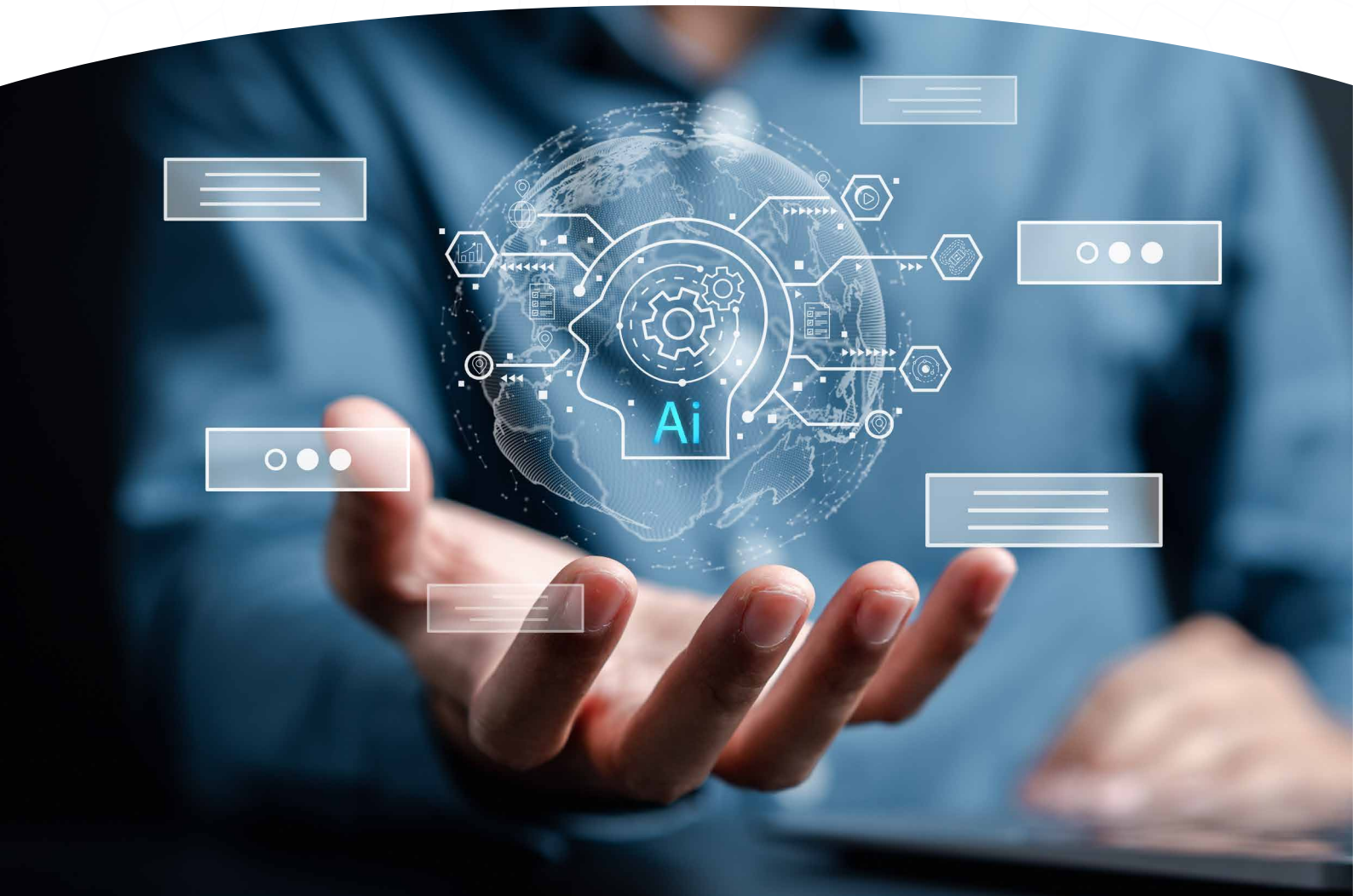
Simple Guide to Maturing Responsible AI

By Sander Ketelaar (Full Orbit)
and Tim Kleinloog (Deeploy)

DEEPLOY

Full Orbit
DRIVE VALUE WITH DATA



If you like to follow the latest developments in AI, you have likely noticed the growing tension between rapid innovations in (generative) AI and regulations. However, this isn't just a conflict between progress and restrictions. It's a critical moment to define how we want to use AI while managing potential systemic risks.

In this whitepaper we do not want to dive into all the complexities of designing proper regulatory and governance frameworks, instead, we want to give you a clear, practical guide to implement Responsible AI. Because when done right, innovation and responsibility go hand in hand.

As an organisation, it is often not straightforward how and where to start or what to focus on. In this guide, we aim to sketch a path towards responsible and mature AI using clear building blocks as a reference.

The state of AI

Artificial Intelligence (AI) has become a ubiquitous technology in our everyday lives, influencing everything from the products we buy to the decisions that are made about us. However, as AI becomes more decisive, it is essential to ensure that it is developed and used responsibly.

According to research by the AIgoSoc consortium (March 2024, 3,946 participants), there are two major challenges in AI adoption in The Netherlands:

- *Low AI literacy* – Many are more familiar with the term ChatGPT than the broader concept of Generative AI.
- *Low trust in AI* – Trust remains a significant issue, with nearly a similar % of people trusting AI to recommend a relevant news article than give an accurate diagnosis for cancer (and both still lower than 45%).

Maturing in responsible AI practices is essential, not only for addressing challenges like these but also for building the confidence needed to drive impactful, trustworthy AI adoption.

The basic principles for Responsible AI

Responsible AI refers to the development and deployment of AI systems that are ethical, transparent, and accountable. Implementing responsible AI is not just the right thing to do, it's also good for business, as it can enhance customer trust and brand reputation.

Rather than being the final step in AI maturity, we believe responsible AI should be integrated into every building block as your organisation grows in AI maturity.

3 guiding principles for Responsible AI

By systematically addressing the foundational principles below, organisations can develop AI systems that are aligned with human values:

1. *There is a clear goal, KPI*
2. *The right data is available for training/evaluation*
3. *Risks are known and can be controlled*

1. *There is a clear goal, KPI*

Imagine AI development like planning a journey. Without a clear destination and way to measure progress, you'll waste resources and potentially head in the wrong direction entirely.

In responsible AI, having a clear goal means:

- Defining exactly what problem the AI system is meant to solve
- Establishing measurable, specific outcomes that define success
- Creating quantifiable metrics that can objectively assess the AI's performance

For instance, when developing an AI for medical diagnosis, a clear goal might be to "Accurately detect early-stage lung cancer with at least 95% precision, while minimizing false positives."

This goal is:

- Specific (lung cancer detection)
- Measurable (95% precision)
- Bounded (early-stage detection)
- Meaningful (reducing diagnostic uncertainty)

2. *The right data is available for training/evaluation*

Data is the foundation of any AI system. Think of data like the ingredients in a recipe, the quality and appropriateness of your ingredients fundamentally determine the final dish's quality.

In responsible AI, having the right data means:

- Trustworthy
- Diverse
- Free from bias

For example, Google's AI for Social Good initiative uses diverse datasets to train models for predicting natural disasters like floods. By ensuring the data is comprehensive and representative, the AI can provide accurate and timely warnings to affected communities.

3. *Risks are known and can be controlled*

AI systems do not contain any magic, they cannot think for themselves (yet), instead, they are complex, sometimes autonomous software systems.

For responsible AI, this means we have to think about potential risks of software, which may include (among others):

- Privacy
- Ethics
- Safety (including security vulnerabilities, see OWASP AI)
- Human oversight
- Transparency

For example, an AI system that tries to optimise the time a user spends in a news app contains incurs in:

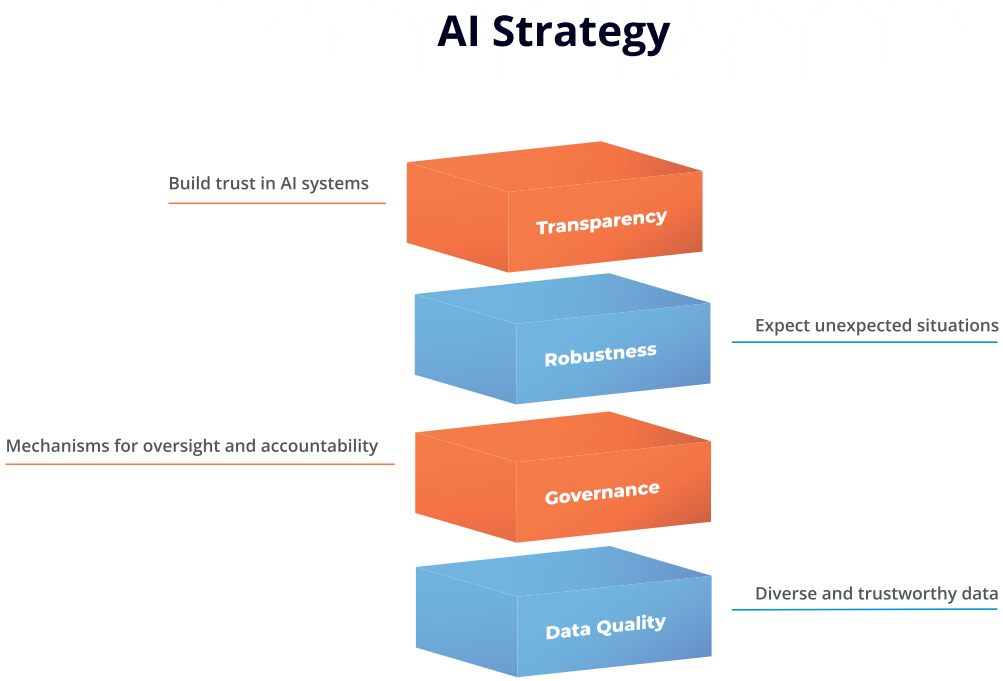
- Privacy risks: measuring news preferences of individuals is private
- Ethical risks: preventing one-sided news coverage may need to be prevented

These three principles aren't isolated, they're deeply interconnected. A clear goal helps you collect the right data. The right data helps you understand and control risks. Risk management ensures your goal remains ethically and practically achievable.

But where to start, what is required and how to grow in maturity?

Maturing in Responsible AI

AI maturity refers to the level at which an organisation has integrated AI capabilities into its operations and decision- making processes. It's not just about having advanced AI technologies, but also about having the right strategies, processes, and cultures in place to support the effective use of AI.
From our experience across different projects, we've identified four key building blocks for organisations to grow and mature in Responsible AI:



In the rest of this guide we will discuss the do's and don'ts for each responsible AI building block and how it impacts AI maturity. Our goal is to help you to implement the right strategy for AI investments in your organisation.

Before starting:

Some of these lessons may seem straightforward, but they can save countless resources and spare you major frustrations.

Do	Don´t
Evaluate if AI is the solution to your problem Define truly measurable assumptions that you can test and validate after development.	**Make using AI the goal** Even if the board has decided that you need to "do something with AI".
Embrace a multidisciplinary approach Mature and responsible AI is not achieved with a one-man/team show. Make sure you know who to involve, and when, and align with your colleagues on risk, security, and compliance.	**See integrating AI as only a technical challenge** Although AI projects certainly include technical challenges, you also want to prevent roadblocks caused by non-technical issues.
Invest in AI literacy Proper understanding of AI also increases support for responsible AI practices.	**Neglect training and education** It is easy to overlook the need to train developers, data scientists, and business leaders on ethical AI development practices, bias detection, and data governance.

Impact on AI maturity: Building a strong business case and involving key stakeholders is not unique to AI projects, but is often forgotten in organisations with a low AI maturity. The challenge here is that teams unfamiliar with AI often struggle to navigate AI-related projects and decisions.

Building Block 1: Data foundation

The foundation of any AI system is the data it is trained on. Therefore, it is crucial to ensure that this data is of high quality and free from bias. This can be achieved by collecting diverse data sets, removing personal information, and applying statistical methods to detect and mitigate bias.

Do	Don´t
Do assess your data foundations Assess the quality and completeness of the data you want to use. Moreover, data can be proprietary, under copyright, or include PII. Always make sure to know and agree on what you can and cannot use.	Skip the data groundwork Data is the fundament of every AI. Garbage will produce garbage as output.
Measure before predicting See AI as a next step of intelligence on top of a KPI that you have been measuring over a longer period.	Without the right data, no magic Don't expect magic things from your AI if you never measured or collected data before about your target variable or important influential variables (e.g., collecting reasons for churn).
Version your data It is essential to relate models to specific versions of data.	Skip data governance Your data team will send gratitude.

Impact on AI maturity: Starting with a basic data foundation allows for flexible ad hoc analysis and modelling. This provides a solid base to build upon, improving reproducibility and enabling automation.

You don't need a complete data foundation for every potential future use case—let it evolve alongside your data needs and organisational maturity. The same principle applies to data quality: focus on it from the beginning and, as you grow, establish an ongoing process to measure, monitor, and ensure its consistency.

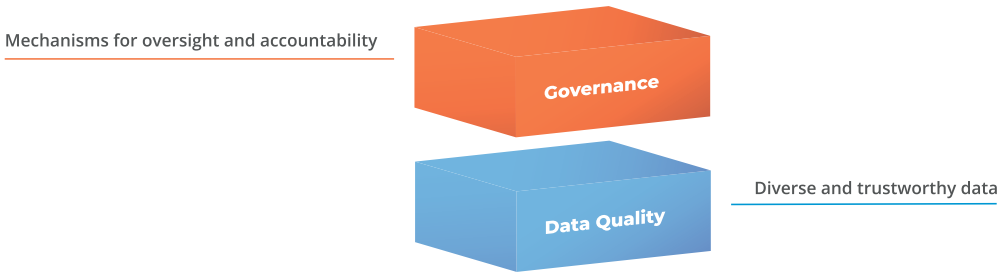
Building Block 2: Governance

Responsible AI requires a strong governance framework that ensures that AI systems are developed and used in an ethical and accountable manner. This is a continuous process that follows the AI lifecycle and includes establishing clear policies and guidelines, ensuring compliance with regulations, and establishing mechanisms for oversight and accountability.

Do	Don´t
Have clear data and model owners and responsibilities defined By establishing clear data and model ownership along with defined responsibilities, organisations can spot data errors or model performance degradation. Ownership ensures that someone is responsible for monitoring, maintaining, and improving these assets over time.	**Violate User Privacy** Avoid collecting data without consent or engaging in unethical data practices, such as scraping personal data without permission.
Implement data privacy and security standards Ensure that AI systems comply with data privacy laws such as GDPR.	**Ignore regulatory changes** Failing to stay up-to-date on regulatory changes can expose your organisation to compliance risks and legal liabilities.
Save logs that have clear audit trails Audit trails provide a transparent, chronological record of activities and changes, offering numerous benefits in terms of accountability, security, compliance, and troubleshooting.	**Treat audit trails as a nice-to-have** Without proper audit trails it is not possible to investigate any incidents that may happen during a model lifecycle, so also for low risk models incidents happen.

Impact on AI maturity: AI adoption should follow a phased approach, scaling use cases alongside governace requirements. Starting with low-risk, high-value applications helps organisations build foundational knowledge and governance structures with fewer compliance demands.

As AI adoption progresses, governance must evolve to handle higher-risk use cases by introducing more stringent oversight, risk management, and ethical guidelines.



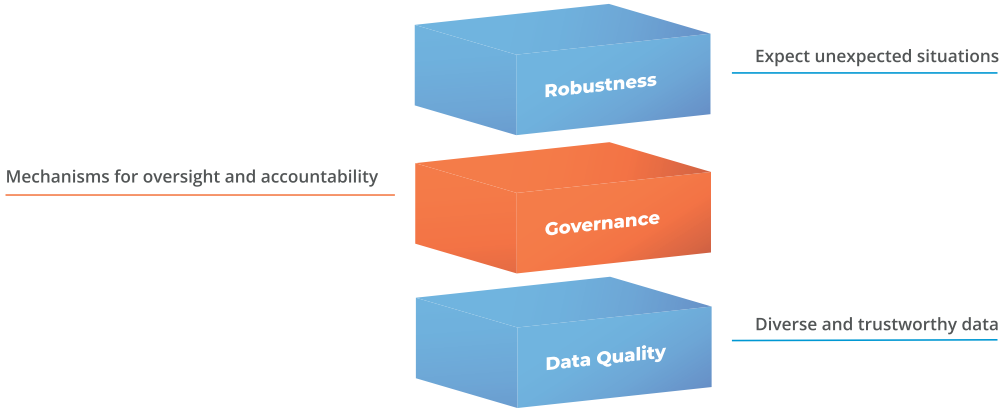
Building Block 3: Robustness

AI systems must perform well under a variety of conditions and should not break down when faced with unexpected situations. This can be achieved by extensively testing and monitoring AI systems to ensure their resilience to changes in the environment. Techniques such as drift monitoring and human feedback loops can alert model owners when to take action to ensure the robustness of the AI system.

Do	Don´t
Test before deploying based on clear goals and requirements Both on initial model deployments and future updates.	Model once, run forever As the world continuously changes, your model should change as well.
Be in control of models with proper monitoring Use technical and statistical monitoring to act in near real-time on anomalies in model behaviour (e.g., model drift)	Only act on technical bugs Keep a close eye on the outcomes of models over time and not only on technical failures.
Be aware of specific AI security risks AI models can be sensitive to attacks, such as adversarial attacks, data poisoning, and privacy leakage.	Only focus on standard software security risks AI models have specific security risks (next to standard security risks). Check for example the OWASP AI

Impact on AI maturity: At the early stages of AI maturity, organisations typically focus on creating models that perform well with the data they were trained on in predictable environments. These models are reliable but often struggle when exposed to variability or unseen data. As AI maturity increases, organisations should invest in more comprehensive testing and validation practices, such as testing for edge cases, handling missing or corrupted data, and evaluating the model's robustness to unexpected inputs.
The next step is to focus on building AI models that can adapt and remain effective when exposed to dynamic, real-world conditions such as model drift and evolving data.

At the highest levels of AI maturity, organisations implement continuous monitoring and feedback loops to ensure AI models are performing optimally even as conditions change. This includes real-time performance tracking, alert systems for potential drift or failure, and feedback mechanisms where users or automated systems can flag issues.



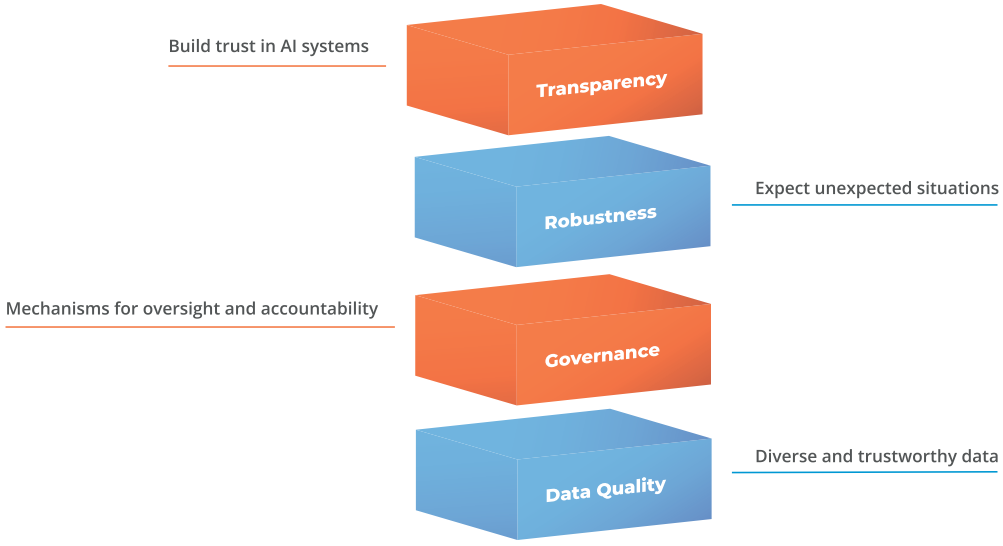
Building Block 4: Transparency

To build trust in AI systems, it is essential to understand how they arrive at their decisions. AI Transparency is the ability of an AI system to provide clear and understandable explanations of its decision-making process. This can be accomplished by using explainable AI (XAI) techniques and translating them into visualisations or instructions that meet the user's needs.

Do	Don´t
Promote transparency and explainability Make AI models and their decision-making processes as transparent and explainable as possible. Ensure that users, customers, and other stakeholders understand how and why decisions are made by AI systems. Also, communicate its capabilities and limitations.	**Treat AI as a "black box"** Failing to comprehend its inner workings can make it harder to identify and mitigate risks.
Human-AI interaction (human oversight) Establish clear roles and responsibilities for AI governance, ensuring human oversight throughout the lifecycle of AI systems—from development to deployment and monitoring.	**Automate everything** Don't replace human judgment entirely with AI in critical decisions (e.g., healthcare diagnoses, legal rulings, or hiring). AI should assist, not replace, human decision-making in high-stakes situations.

Impact on AI maturity: As organisations advance in their AI maturity, they shift from simple technical documentation (like model cards) to more comprehensive, accessible, and accountable frameworks for transparency.

In early stages, organisations often offer basic technical documentation of how AI models work, focusing on the algorithms used. However, documents are typically aimed at technical experts and lack clarity for non-technical stakeholders.
As maturity grows the focus shifts towards making model decisions more transparent to end-users by providing understandable explanations of the key features that influence outcomes. This allows users to see why certain decisions were made.
At the highest level of maturity, organisations not only make models fully explainable but also implement accountability frameworks that allow users and regulators to audit and challenge AI decisions.



Shaping the path to Responsible AI

Building responsible AI systems is a continuous journey, and using these building blocks as a framework can help fine-tune your strategy for long-term success.

By matching your organisation's needs to its current level of development and using a step-by-step growth approach, the costs, rules, and requirements for building and running AI responsibly and safely become easier to handle.

For those looking ahead, following these building blocks will not only strengthen your AI maturity but also position you to easily comply with regulations like the EU AI Act.

Just ensure all building blocks are solidly in place before scaling up your AI initiatives to avoid potential risks down the line.

About Deeploy

Deeploy was founded with the intent to pioneer and innovate in the field of control, transparency, and explainability of AI models, especially when it comes to high-risk applications. Deeploy's focus lies on building blocks Governance, Robustness, and Transparency.

About Full Orbit

Full Orbit is a professional services company specialized in data management and AI, providing end-to-end solutions that encompass all four key building blocks mentioned in this blog. Leveraging their expertise in both data management and AI, Full Orbit is equipped to develop, deploy, and maintain AI systems efficiently and effectively.