BOSCH

Responsible AI

Technology for Trust | Trust in Technology

BOSCH

Table of Contents

Introduction

Today, we stand at a pivotal moment in the evolution of AI. Two critical developments are unfolding in parallel: on one hand, tech giants are equipping AI developers with cutting-edge tools to build, test, and refine AI solutions responsibly. On the other, governments are actively working to establish regulatory frameworks, policies, and laws to guide the adoption of AI. While these two tracks currently operate largely in isolation—leading to ambiguity around what constitutes responsible AI—they are on a path to converge. Soon, we will live in a world where AI governance is seamlessly embedded into enterprise processes and societal systems alike.

The democratization of AI has made powerful tools accessible to virtually everyone. Whether through open-source models, cloud-based platforms, or specialized software, individuals and organizations can now readily harness AI technologies. This has sparked a wave of innovation, but it also brings significant risks—from biased outputs and privacy violations to the spread of AI-generated misinformation.

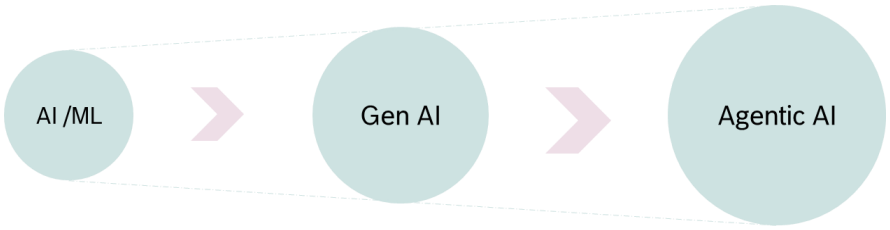
Trust is built on quality, and trust drives adoption—ultimately fueling business success. Creating a lasting AI success story requires more than innovation and new business opportunities; it demands that we earn society's confidence in AI's benefits. We are not just building AI—we are building trust in AI.

From diagnosing diseases in healthcare to optimizing supply chains in Manufacturing, streamlining hiring in HR, and personalizing experiences in smart homes, AI's ability to learn, predict, and automate is unlocking unprecedented efficiency and innovation. But as these systems take on more decision-making roles, ensuring they act responsibly becomes critical. In healthcare, responsible AI means prioritizing patient safety and equity, not just statistical accuracy. In manufacturing, it involves transparent algorithms that don't compromise worker safety for efficiency. In HR, it ensures hiring tools are free from bias and respect candidate privacy. In home automation, it means safeguarding user data and maintaining control in human hands. Responsible AI is about aligning innovation with ethical standards to build systems that are not only smart, but trustworthy.

The risk matrix - from Classical to GenAI to Agentic AI

AI, Generative AI (GenAI), and Agentic AI represent the three key stages in the evolution of intelligent systems. AI enhances decision-making, GenAI enables seamless content creation, and Agentic AI empowers systems to act independently toward defined goals.

More recently, AI conversations have shifted from "*How does this work?*" to "*How quickly can we use it?*" Unlike traditional machine learning and deep learning models that required expert handling, GenAI is designed for accessibility, shifting the focus from algorithmic transparency to real-world value creation. Agentic AI takes this further by enabling systems to make autonomous decisions and take independent actions to achieve complex goals.
With this evolution, the nature of AI risks is shifting—from controlled errors in traditional models to unpredictable outcomes and ethical challenges in more autonomous systems.



AI Type	Primary Risks	Vulnerabilities	Responsible AI Focus
Traditional ML	Bias in data, lack of transparency	Decision-making	Fairness, Explainability
Generative AI	Misinformation, IP issues, harmful outputs	Deepfakes, Misinformation	Safety, accuracy, responsible data use
Agentic AI	Unpredictable behavior, loss of control, value misalignment	Unpredictable security vulnerabilities	Oversight, alignment with human intent, control

Prioritizing speed over scrutiny, GenAI and Agentic AI have moved from research labs to boardrooms at an unprecedented pace. Unlike earlier AI models, which were tested in controlled environments, GenAI is being openly deployed in consumer-facing applications. Users and enterprises are captivated by its results—compelling text, realistic images, and automation—often without questioning its underlying decision-making process. This differs from AI in regulated industries like finance and healthcare, where audits and explainability were essential for adoption.

AI models have evolved from thousands to trillions of parameters, making GenAI inherently dynamic and unpredictable. Its responses can vary, and its ability to engage in human-like conversations at scale increases the risk of misinformation, bias, IP infringement, and operational errors. Agentic AI introduces a new level of complexity by autonomously navigating environments and making goal-driven decisions, increasing the need for oversight to manage risks related to control and alignment.

Without a strong RAI foundation, we risk losing public trust, facing regulatory backlash, and ultimately stalling innovation.

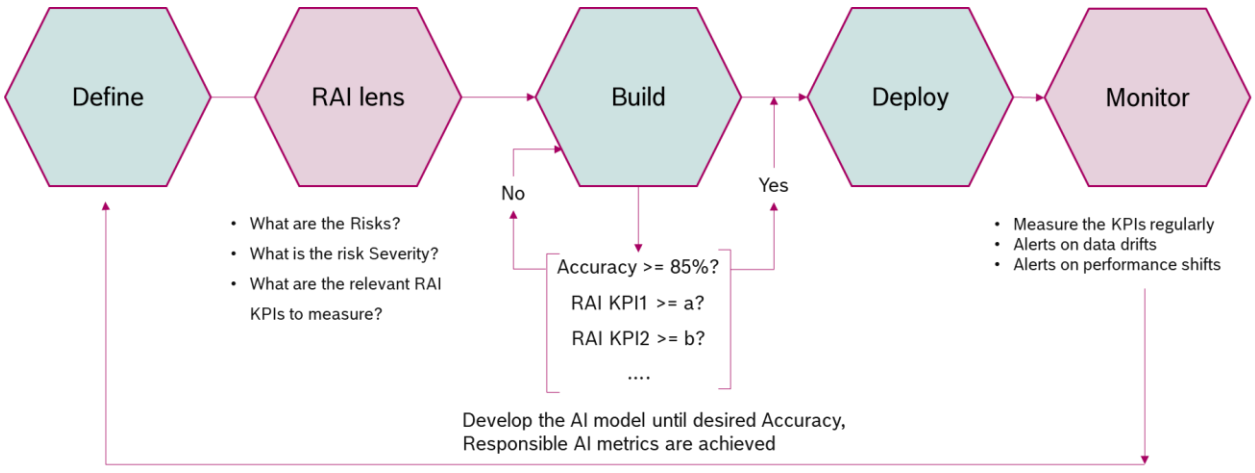
Re-define the process of creating AI-based solutions

We need to rethink how we build AI solutions. Traditionally, the process followed a simple flow: define the use case, develop the model to meet accuracy targets, then deploy it as part of a product or service. But real-world AI behavior often differs from controlled tests, leading to challenges such as lack of transparency and unintended harm. Some real-world failures of AI include:

- Discriminatory hiring practices resulting in costly lawsuits (e.g., iTutorGroup settlement)
- Biased algorithms limiting healthcare access for millions
- Chatbots providing legally binding or harmful advice (e.g., Chevrolet and NEDA cases)
- Misinformation and "hallucination" from AI-driven customer service

In order to mitigate the probability of such risks, we need to re-imagine our process of creating AI-based solutions:

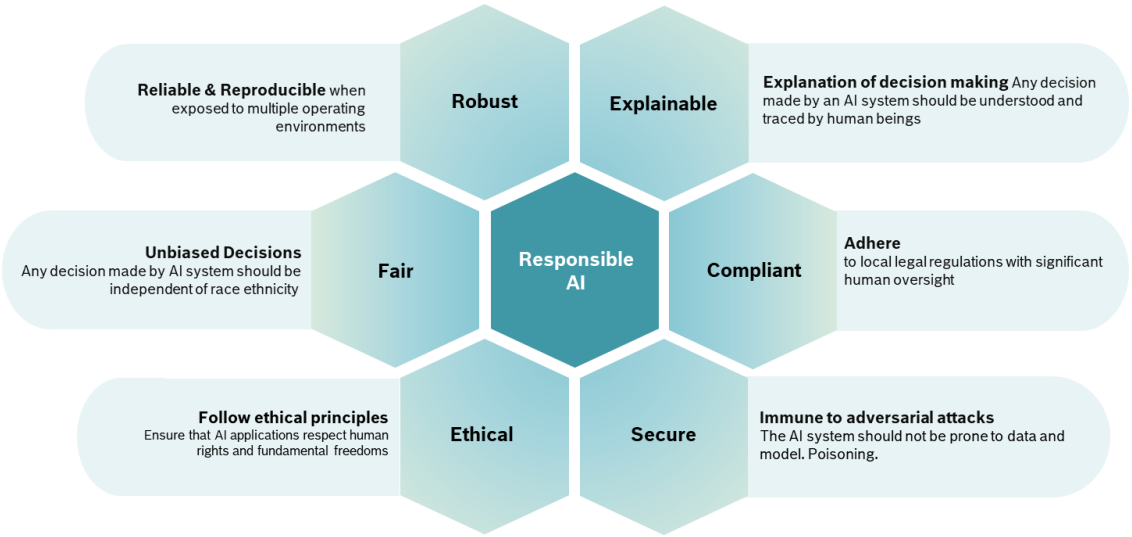
1. **Define** - Begin by identifying the AI use case and setting clear goals for what the system should achieve.
2. **RAI Lens** - Assess all possible risks by AI. Understand the severity of those risks and determine the Responsible AI KPIs that will entrust responsible behavior and must be tracked throughout the lifecycle.
3. **Build** - Develop the AI model with a dual focus—achieving target accuracy and meeting Responsible AI benchmarks. The model iterates until both performance and responsibility metrics are satisfied.
4. **Deploy** - Once thresholds are met, deploy the model as part of the product or service. At this stage, the system is not only functional but also responsibly aligned.
5. **Monitor** - Continuously measure the performance and Responsible AI KPIs, track for data drift or performance shifts, and set up alerts to detect and respond to any deviations in real time.



Example: A car manufacturer developing an AI-based lane detection system begins by defining its goal—accurate lane recognition for driver assistance. Through a Responsible AI lens, the team identifies risks like failure in poor weather or on worn-out roads and sets KPIs for robustness and fairness. During development, the model is trained and refined to meet both accuracy targets and RAI benchmarks. Once validated, it's deployed into the vehicle's software. Post-deployment, the system is continuously monitored for performance, data drift, and safety deviations, ensuring it stays reliable and responsible in real-world conditions.

The pillars of responsibility

Based on the key principles from NIST, EU AI Act, and OECD recommendations, the suggested core principles of Responsible are:



Fair - AI should treat all individuals equitably. This means proactively identifying and eliminating biases in data, algorithms, and outcomes to avoid discrimination in areas like hiring, lending, or healthcare.

Robust - AI systems must be resilient and dependable under varied conditions. Whether facing noisy data, edge cases, or real-world uncertainty, models should perform reliably without compromising accuracy or safety.

Explainable - Users should be able to understand how an AI system arrives at its decisions. Transparent models improve user trust, support accountability, and are critical for sectors requiring auditability, like finance or healthcare.

Compliant - AI must follow evolving regulations and legal frameworks. This includes data privacy, human oversight, and region-specific requirements, ensuring systems operate within lawful and ethical boundaries.

Secure - AI must be protected against threats such as data poisoning, adversarial attacks, and unauthorized access. A secure AI system safeguards both users and organizations from manipulation and harm.

Ethical - AI should reflect and respect ethical principles, including fairness, dignity, and autonomy. It should enhance human well-being and avoid causing social or psychological harm.

Responsible AI metrics for Classical, GenAI, Agentic AI

Responsible AI metrices for Classical AI/ML

Explainability- Techniques to interpret and explain AI decisions:

- **Model-Agnostic Explainers:** SHAP and LIME provide local explanations for individual predictions.
- **Gradient-Based Methods:** Use model gradients to highlight important features influencing outputs.

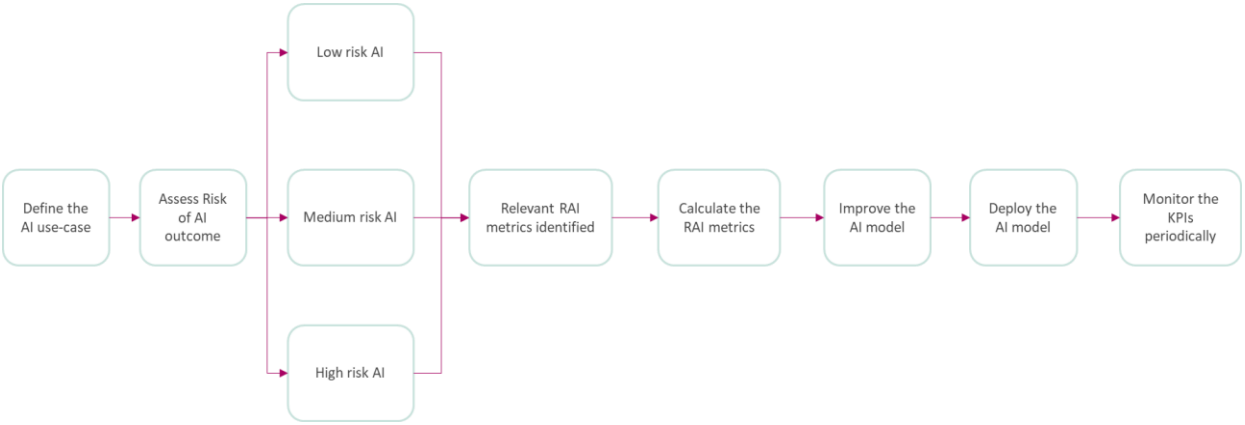
Fairness - Ensuring equitable AI outcomes across diverse populations:

- **Parity Measures:** Demographic, impact, predictive, precision, and statistical parity assess balanced treatment and error rates across groups.

Robustness

Maintaining reliable AI performance under challenging conditions:

- **Adversarial Training & Counterfactual Analysis:** Strengthen resistance to malicious or borderline inputs.
- **Randomized Smoothing & Outlier Detection:** Improve stability and identify anomalies.



Responsible AI metrices for GenAI

Language Quality

Evaluating fluency, coherence, and naturalness:

- **Perplexity and BLEU:** Automated metrics for language modeling and similarity to reference texts.
- **Human Evaluation:** Expert assessment of quality.

Relevance & Context Understanding

Ensuring outputs are contextually appropriate and meaningful:

- **ROUGE and Semantic Similarity:** Measure content overlap and meaning alignment.
- **On-Topic Rate:** Percentage of relevant generated responses.

Efficiency & Performance

Balancing speed and resource use:

- **Latency, Throughput, and Token Efficiency:** Metrics for response time, request handling capacity, and computational cost.

Safety & Robustness

Minimizing harmful or biased outputs:

- **Error Rate, Bias & Fairness, Toxicity Scores:** Monitor and mitigate unsafe content.
- **Adversarial Robustness:** Protect against malicious manipulations.

Retrieval-Augmented Generation (RAG) Metrics

For models combining retrieval and generation:

- **Faithfulness, Relevance, Coherence, and Factuality:** Ensure accurate, contextually precise, and truthful outputs.
- **Noise Sensitivity and Multimodal Evaluation:** Assess resilience to irrelevant data and performance across data types.

Text-to-SQL Metrics

Assessing natural language to SQL query translation:

- **Query Correctness and Semantic Equivalence:** Measure functional accuracy and meaning preservation.

Responsible AI metrices for Agentic AI

Performance

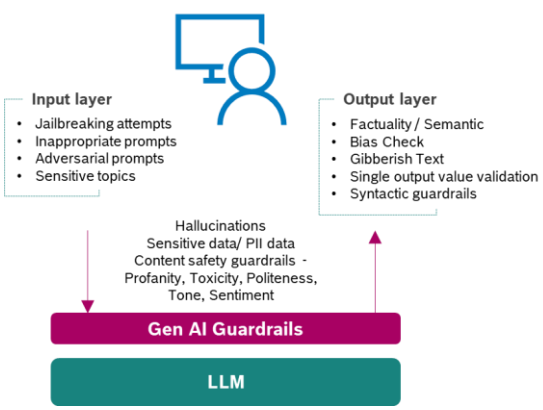
- **Task Success, Latency, Throughput, Goal Alignment, Resource Utilization, and Accuracy:** Key indicators of task effectiveness and operational efficiency.

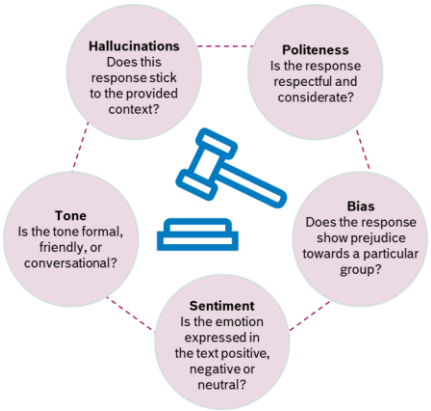
Security

- **Resistance to Adversarial Attacks, Data Leakage, Vulnerabilities, Traceability, and Prompt Injection:** Safeguarding system integrity.

Robustness

- **Error Recovery, Out-of-Distribution Performance, Stress Testing, Graceful Degradation, Adaptability, and Fault Tolerance:** Ensuring resilience in dynamic environments.





Governance and Guardrails

Robust guardrails are essential for ensuring AI systems operate safely and ethically. The LLM-as-a-Judge framework leverages large language models to continuously evaluate AI outputs in real-time, detect harmful or unintended responses, and enforce adherence to ethical guidelines and organizational standards. This proactive oversight helps maintain trust, accountability, and alignment throughout the AI lifecycle.

Sure.AI – Suite of Responsible AI

With Bosch Global Software Technologies (BGSW spearheading AI and Data Science across various verticals—leveraging technologies such as computer vision, deep learning, and generative AI in automotive, manufacturing, and healthcare domains—a strategic move towards establishing in-house RAI processes and tools appears to be a prudent decision. Along with positioning the brand towards a Responsible engineering company, such a step will enhance customer satisfaction by building trust in our AI-based products and services, ultimately leading to increased business opportunities.

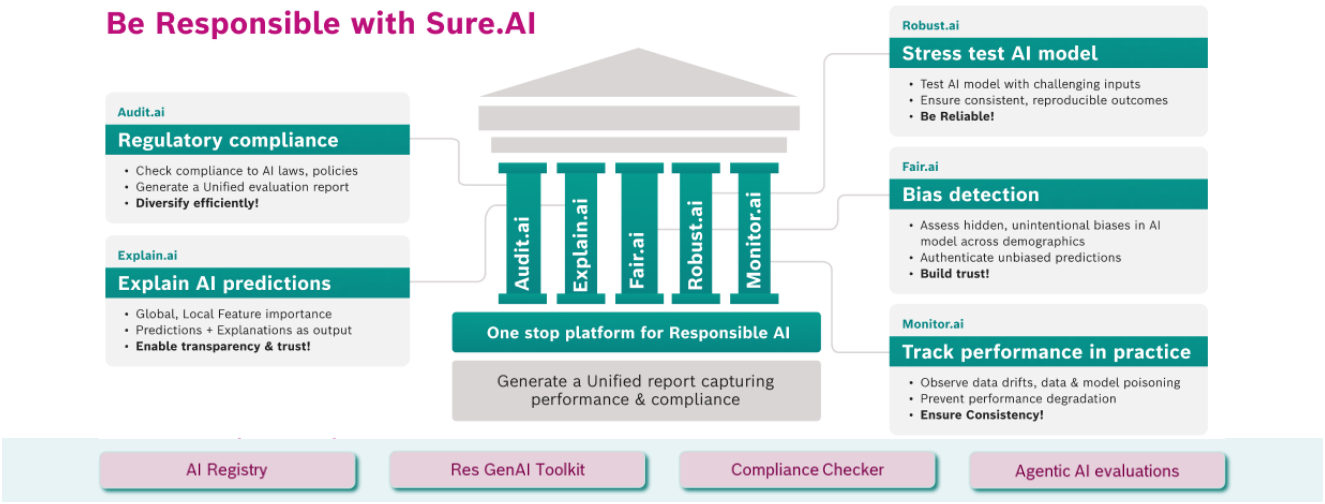
We are developing an end-to-end RAI workflow platform: Suite for Responsible AI (Sure.AI), inspired by the trustworthy AI framework curated by HLEG formed by the European Commission and ACM FAccT. The platform facilitates organization-wide visibility of AI compliance and seamlessly marries the legal and technical aspects of RAI with insightful dashboards, risk assessments, compliance reports, and technical evaluations of AI models—all culminating in a Unified Report of Responsible AI.

The platform is built on five key pillars:

- **Audit.AI** assists regulatory compliance by offering risk assessment, compliance checklists, and audit trail completeness.
- **Fair.AI** identifies bias detection in AI predictions using fairness metrics such as Equal Opportunity Difference, Disparate Impact Ratio, among other KPIs.
- **Explain.AI** provides local and global model explainability, delivering human-understandable explanations and traceability to elucidate model decisions.
- **Robust.AI** confirms reliable and reproducible outcomes through rigorous stress testing, using metrics like Adversarial Robustness and Performance Consistency.
- **Monitor.AI** detects performance degradation and data drifts to ensure data integrity and measures anomalies post-production.

A comprehensive health report of the AI model is delivered as a thorough audit of performance, compliance, and ethical adherence. The report identifies areas for enhancement with actionable recommendations and supports informed decision-making.

Assenting to such a platform within Bosch supports our AI initiatives every step of the way. It improves transparency, gives our products a competitive edge, increases the likelihood of certifications, and makes us future-ready for RAI regulations. Entering late to the market, on the other hand, risks losing competitive advantage and facing delays in complying with swiftly evolving RAI regulations.



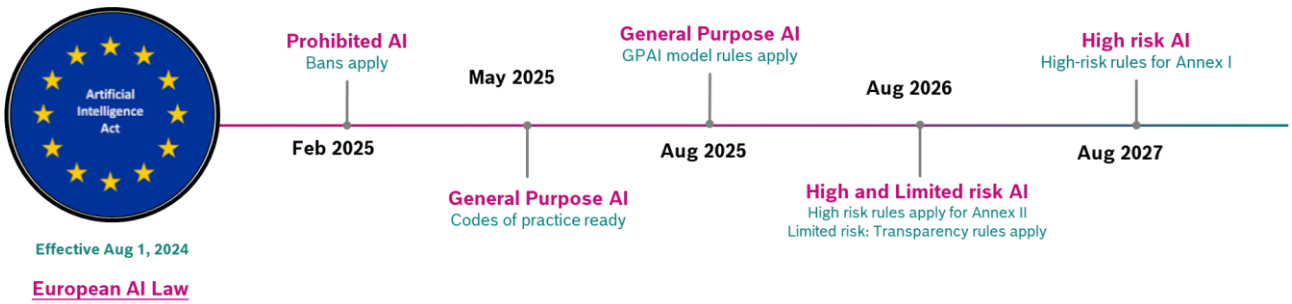
AI Registry: Centralized catalog for tracking and managing AI models and assets.

Res GenAI Toolkit: A set of tools designed to develop, evaluate, and optimize generative AI systems responsibly.

Agentic AI Evaluations: Framework for measuring performance, security, and robustness of autonomous AI agents.

The EU AI Act

The **EU AI Act** is the European Union's landmark regulatory framework aimed at ensuring the safe and ethical development and deployment of artificial intelligence. It introduces a **risk-based approach**, classifying AI systems into categories such as minimal, limited, high, and unacceptable risk, with **stricter requirements**—including transparency, human oversight, and robustness—for high-risk applications. The Act promotes innovation while safeguarding fundamental rights, setting a global precedent for AI governance and accountability. Once fully enforced, it will apply to any AI system impacting EU citizens, regardless of where the provider is based.



Shape the Future - Be Responsible!

The journey toward Responsible AI is not a one-time effort—it is an evolving commitment to innovation, governance, and trust. As AI continues to shape industries and societies, organizations must proactively embed responsible practices into every stage of the AI lifecycle.

By bridging the gap between technical advancements and regulatory integrity, we are not just responding to challenges—we are defining the future of AI governance. The road ahead requires collaboration, transparency, and continuous iteration. With the right frameworks, tools, and a shared commitment to ethical AI, we can unlock AI's full potential while ensuring its benefits reach everyone.

The time to act is now. Responsible AI is not just a safeguard—it is the foundation for sustainable innovation, competitive advantage, and long-term success in the AI-driven world.