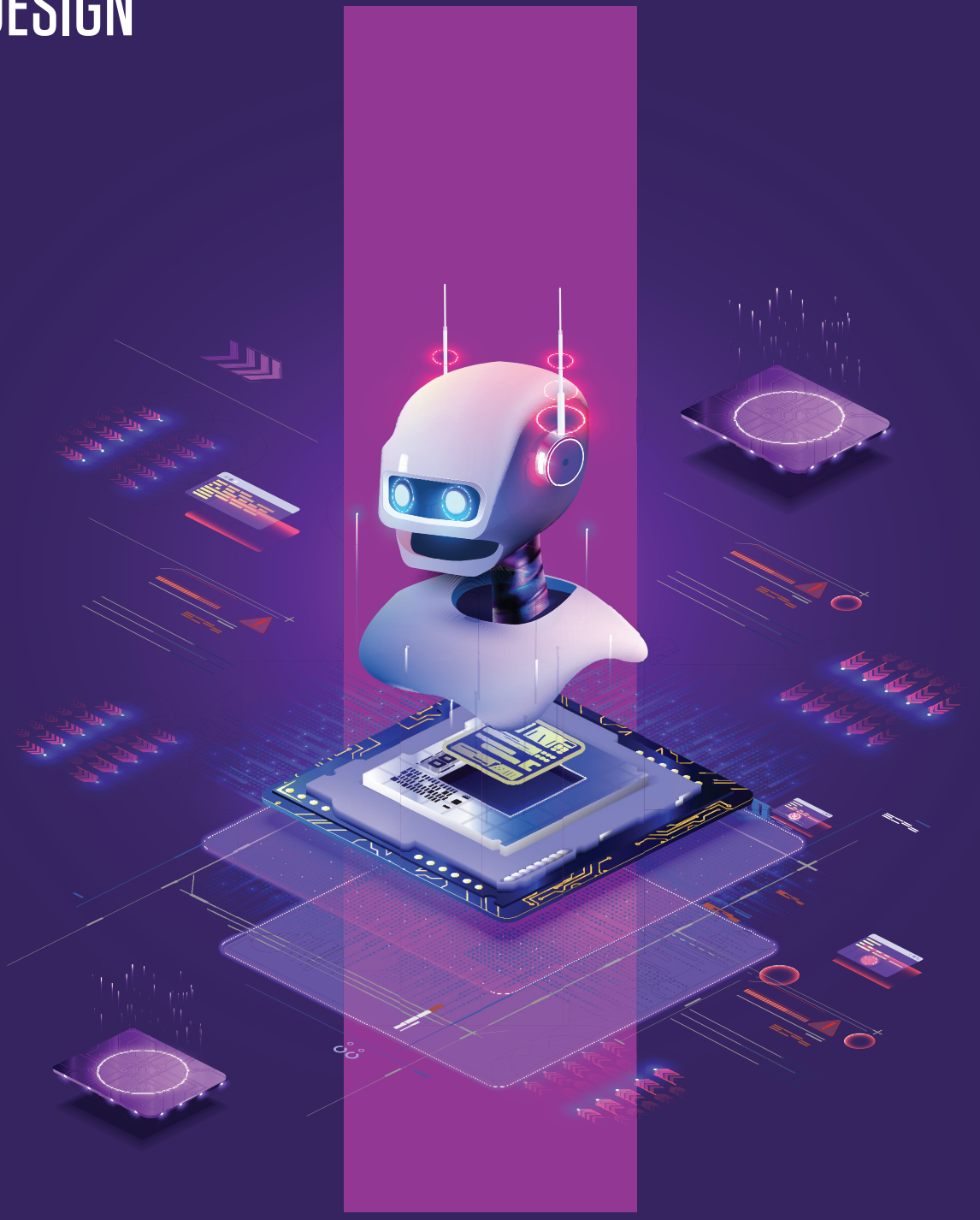
# TECH NAVIGATOR: RESPONSIBLE AI BY DESIGN

Infosys®
Navigate your next



Infosys® | Knowledge Institute

# Contents

AI is now part of our lives, from building products for businesses and consumers to scanning resumes for recruitment and managing remote workers, and from drug discovery to diagnostics. Yet security and governance are only just catching up with the explosion of AI functions.

We are now in the third wave of AI evolution (H3). The first wave (H1) was driven by machine learning, and the second (H2) by deep learning. This third wave is led by foundation models trained on broad data that use self-supervision at scale and which can be adapted to perform a wide range of tasks. In short, these are the models that underpin generative AI, including the large language models that drive tools such as Google's Bard and OpenAI's ChatGPT.

This third wave of AI evolution has brought a new range of concerns to already well-articulated worries about transparency, explainability, human oversight, compliance, and continuous improvement. Generative AI creates copyright concerns around the billions of parameters used to train large models, as well as fears about perpetuating bias and disadvantage, malicious use of AI-generated content, and limited access to foundation models and training data or weights. This limits our ability to address the underlying limitations of these models. We neither choose the data these models are trained on, nor provide human preference data used in reinforcement learning with human feedback (RLHF[1]).

What we can control is a careful evaluation of the model outputs to actively search for biases or mistakes that may arise.

## Appropriate governance

Therefore, appropriate governance is a key plank to becoming an AI-first organization.

At the heart of ethical AI is the concept of "responsible by design". This is already familiar to cybersecurity professionals, who adopt this framework to create and deploy security products and policies. The aim is to bake in security — and now AI ethics — at every step of the process.

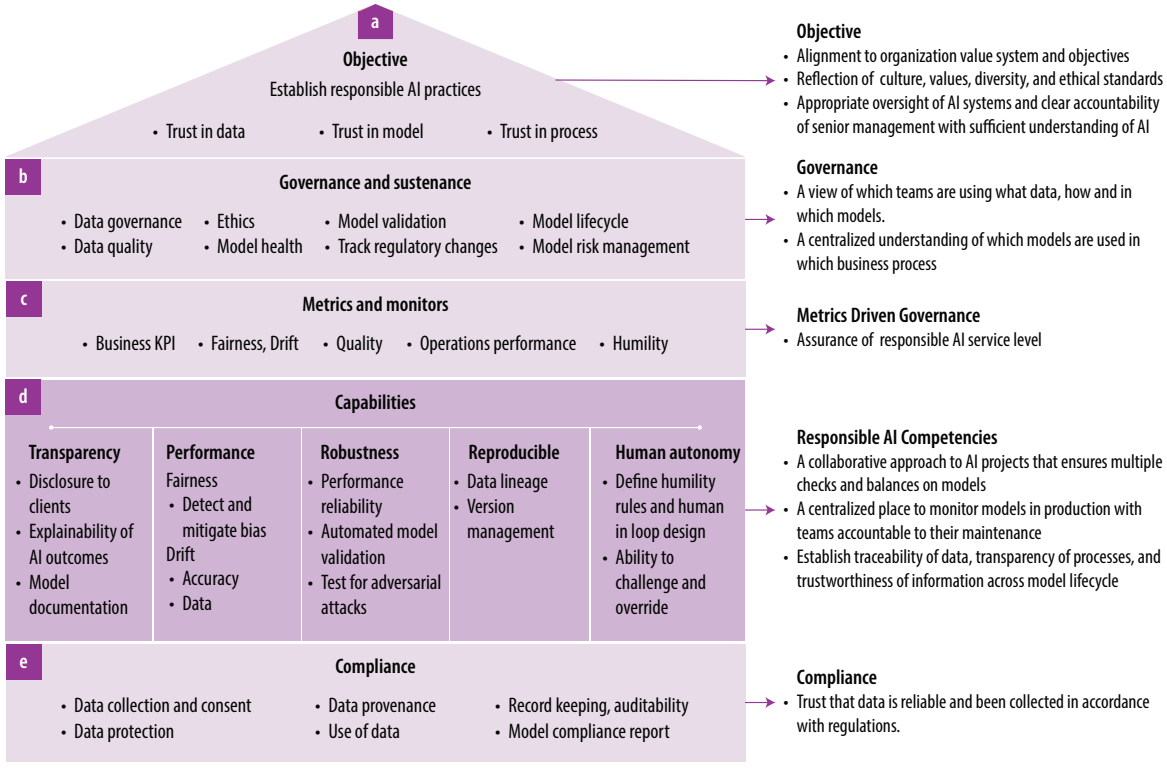
The principles to guide responsible-by-design AI include:

- Human oversight and governance at every stage.
- Constant auditing of processes and products for fairness, inclusiveness, and prevention of harm.
- Transparent and explainable AI that is reliable, safe, secure, and private at each stage.

How should organizations implement these principles?

We have identified five responsible-by-design building blocks: objective, governance, metrics, capabilities, and compliance (Figure 1). Included in these building blocks are

Figure 1. Five responsible-by-design building blocks




Source: Infosys

a focus on aligning with the organization's wider objectives, value systems and diversity standards, as well as details of governance such as assuring data quality, validating the model and tracking regulatory changes.

For human oversight and governance, organizations should identify a diverse range of stakeholders at the outset and engage with them regularly. A wide range of voices that are heard and acted upon will raise potential issues early, and their warnings should mitigate potential harm.

It's vital for organizations to assign responsibility for human oversight, making sure that business sponsors from the outset understand the intended purpose of the AI. This includes compliance teams making sure the AI meets regulatory requirements and does not perpetuate biases. It also includes Ops and IT teams making sure a model can be explained to regulators and is continuously monitored for accuracy.

## Continuous auditing

Auditing requires a cross-functional approach: it cannot be left to solely to product and tech teams. It means working with legal teams, with data protection and with cybersecurity teams to review implementation across the business and to consider how AI products and processes must comply with laws and regulations.

Continuous auditing also means continuous feedback so that engineering teams respond to problems that arise, such as bias in AI decision-making and hallucination in generative AI chatbots. A moderation function can continually review the AI's output and flag problems for fixing.

For foundation models and large language models in Horizon 3, an external control system can also mitigate unintended hallucinations. Currently a few external control platforms facilitate flow between these large language models alongside external sources to augment the model's knowledge, reasoning, and actuation. External controls also create the necessary safeguards for responsible AI design. This is why prompting is important to promote responsible AI design and development practices.

AI products and processes must be transparent and their outputs explainable. This means that the algorithms and inputs that led to a decision or other output can be checked and understood by humans in the business, and by customers and users outside the business. This increases trust in AI systems, which is a foundational prerequisite for deploying advanced AI systems.

## Trust: the next big challenge

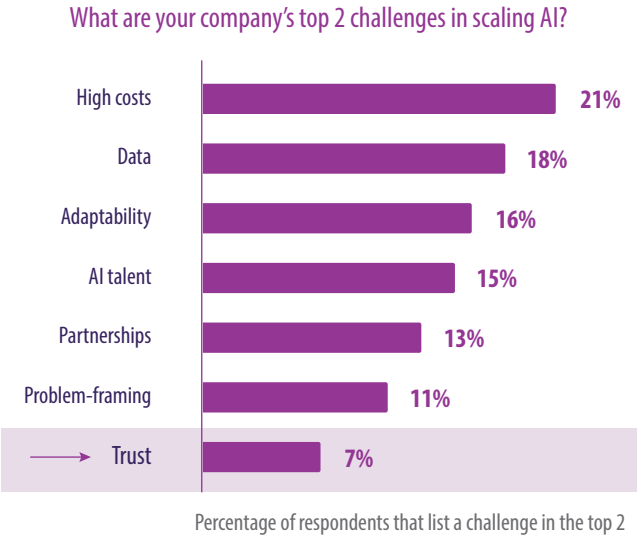
As we wrote in Data + AI Radar[2], trust is the next big challenge in implementing AI systems. When AI systems are deployed at scale, "trust and responsible AI systems become a major issue," says Bonnie Holub, a data science leader with Infosys Consulting. "We see trust [and responsible AI] as crucial parts of the nonfinancial governance issues investors are demanding from companies," she adds. However, despite its importance, executives rated trust as a low concern when surveyed for Data + AI Radar (Figure 2).

Reliability, safety, security, and privacy occur when AI tools, products, services, and processes are built to robust standards and operate consistently, in the way they were originally designed. They should also continue to work as designed under unexpected conditions, and regular testing for reliability must be a part of the design, implementation, and maintenance processes.

An organization can only consider itself an AI-first entity if it has embedded responsible-by-design principles into every corner of its business and work. This applies to those building these tools and those using and deploying them.

Only then can an organization be truly ready to be a part of this transformative and exciting landscape.

Figure 2. Despite its importance, executives rated trust as the lowest concern




Source: Data + AI Radar, Infosys

## References

1. Illustrating reinforcement learning from human feedback (RLHF), Nathan Lambert, Louis Castricato, Leandro von Werra, and Alex Havrilla, Dec. 9, 2022, Hugging Face.
2. Data + AI Radar 2022: Making AI real, Chad Watt, 2022, Infosys Knowledge Institute.

## Authors

**Rajeshwari Ganesan**
Distinguished technologist, Infosys

**Rajeev Nayar**
CTO of data and AI, Infosys

**Kamalkumar Rathinasamy**
Distinguished technologist, Infosys

**Rafee Tarafdar**
CTO, Infosys

**Kate Bevan**
Infosys Knowledge Institute

**Harry Keir Hughes**
Infosys Knowledge Institute

# About Infosys Knowledge Institute

The Infosys Knowledge Institute helps industry leaders develop a deeper understanding of business and technology trends through compelling thought leadership. Our researchers and subject matter experts provide a fact base that aids decision-making on critical business and technology issues.

To view our research, visit Infosys Knowledge Institute at infosys.com/IKI or email us at iki@infosys.com.

For more information, contact askus@infosys.com