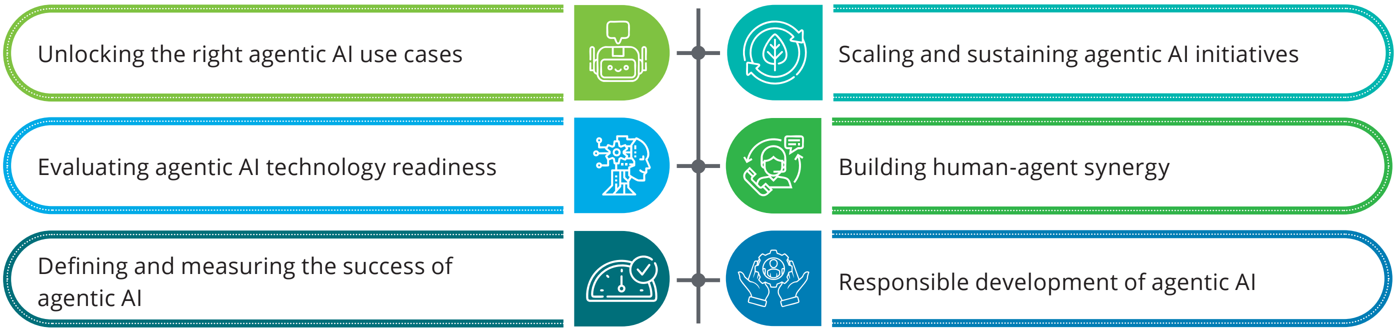
Responsible agentic AI by design

December 2025

Table of contents

Introduction

Agentic AI is shifting the enterprise narrative from task automation to goal-oriented autonomy. Where GenAI systems generated content and assisted knowledge workers, agentic AI can plan, act and collaborate across systems, initiating workflows, orchestrating other agents and driving decision-making with minimal human intervention.[1] Every step forward in capability demands a step up in responsibility. Before organisations scale agentic AI, they must consider the following six key criteria:



Unlocking the right agentic AI use cases

Evaluating agentic AI technology readiness

Defining and measuring the success of agentic AI

Scaling and sustaining agentic AI initiatives

Building human-agent synergy

Responsible development of agentic AI

It is crucial to revisit these six considerations through a lens of responsibility and guardrails, outline the risk landscape specific to agentic AI and translate responsible AI principles into concrete design patterns, controls and governance mechanisms suitable for enterprise deployment. These should align with leading frameworks such as the National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (NIST AI RMF)[2] and the Organisation for Economic Co-operation and Development (OECD AI) Principles.[3,4]

Re-viewing the six questions through a responsibility lens

Unlocking the right agentic AI use cases

This chapter introduced a structured approach to qualify use cases for agentic AI, focusing on reasoning intensity, autonomy, multistep workflows, cyclicity and learning potential.[5]

From a responsibility perspective, this selection step marks the point where risk concentration begins. Processes that are:

- High in stakeholder impact (e.g., credit decisions, clinical routing, compliance checks)
- High in autonomy and escalation authority
- Embedded in regulatory or customer-facing contexts

should be subject to stricter eligibility criteria and additional guardrails (e.g., mandatory human-in-the-loop or human-on-the-loop).

Key questions to ask:

- Does this process require explanations and audit trails for external regulators or internal assurance?
- Could an error propagate into irreversible harm (financial, health, safety, reputational) despite existing policies and risk controls that the agent must respect?

A process should only move from "GenAI-assisted' to 'agentic" when the answers to key responsibility questions are fully understood and validated.

Evaluating agentic AI technology readiness

This chapter outlines the core components of an agentic AI ecosystem, such as foundation models, knowledge and data layers, orchestration and agent frameworks, tooling and integration and DevOps/AgentOps.[6]

A responsibility lens adds three further requirements that echo NIST AI RMF's Map–Measure–Manage functions and its emphasis on secure, explainable and governable AI. These requirements are:

- **Data governance by design:** Data lineage, minimisation, quality checks and access controls must be embedded into data lakes and vector stores that feed agents.
- **Secure connectivity and zero-trust principles:** API gateways, secrets management and threat protection for agents that interact with systems, trigger transactions or modify records.
- **Observability and auditability:** Logs, traces and evaluation tools tailored for LLMs and agents to monitor reliability, drift and policy violations.

Key questions to ask:

- Can we trace which agent did what and with which inputs at any time?
- Do we have kill-switch and rollback mechanisms for agents that malfunction or behave unexpectedly?
- Are our AgentOps pipelines designed to include risk and ethics checks in addition to performance benchmarks?

Defining and measuring the success of agentic AI

Agentic initiatives often start with productivity narratives such as reduced handle time, faster response and lower costs. Over time, success must be reframed more holistically, in line with international guidance on trustworthy AI. This broader view of impact includes three dimensions:

- **Business outcomes:** Revenue uplift, risk reduction, resilience
- **Human outcomes:** Job enrichment, error reduction, well-being
- **Trust outcomes:** Fewer incidents, customer satisfaction, regulator confidence

This chapter suggested evaluating the impact across market opportunity, strategic importance, right to play and economics of the agentic solution.[7]

For responsible deployment, organisations should extend by introducing **"responsibility KPIs",** such as

- Bias and fairness metrics across segments
- Rate of overridden or escalated decisions
- Incidents of policy violations or unsafe recommendations
- Explainability coverage (e.g., Percentage of decisions for which explanations are available and used)

Key questions to ask:

- What mechanisms will identify when the agent is optimising for an unintended proxy metric?
- Which conditions qualify as red lines requiring us to slow, pause or roll back deployment?

Beyond the pilot: Approach to building, scaling and sustaining agentic AI initiatives

This chapter emphasized the need for a roadmap and prioritisation framework, balancing impact, ease of implementation and differentiability.[8]

A responsible scale-out roadmap includes:

- **Phased deployment:** Starting with low-risk, low-complexity use cases, then expanding to high-impact, higher-risk domains as governance matures.
- **Central agentic AI governance:** A cross-functional group (business, risk, legal, technology, HR) that owns standards, patterns and approvals, reflecting the "Govern" function in NIST AI RMF. [7]
- **Reusable guardrail patterns:** Standardised approaches for logging, human-in-the-loop, escalation and monitoring that every new agent must implement.

Key questions to ask:

- Are we avoiding one-off pilots and building a reusable, governed stack?
- Do we have a risk-based tiering of agentic use cases with corresponding controls?

Building human-agent synergy: Equipping and empowering employees to thrive alongside AI agents

This chapter highlights that as agentic AI takes over routine tasks, the workforce can focus more towards meaningful, value-driven roles.[9]

Responsible deployment requires intentional design of:

- **New roles:** From "operator" to orchestrator, supervisor and auditor of agents
- **Capability building:** Prompts literacy, effective oversight and proactive risk awareness for frontline employees
- **Incentives and performance management:** Compensation and incentives tied to output and responsible AI governance, reflecting OECD principles on inclusive growth and human-centred values.

Key questions to ask:

- Are managers and teams equipped to critically question, challenge, and override agent decisions when necessary?
- What measures address the psychological and ethical burden on human supervisors responsible for approving or vetoing automated actions?

Responsible agentic AI by design

This chapter consolidates earlier discussions and examines the operationalisation of responsibility across the design, deployment, and operation of agentic AI.

The remainder of this chapter focuses on the following question.

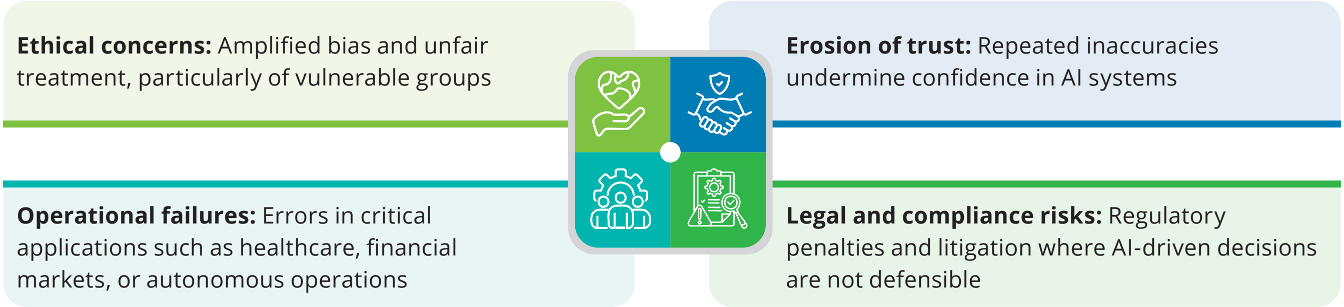
What frameworks and practices support the responsible design, deployment and operation of agentic AI?

The risk landscape: Why agentic AI needs stronger guardrails

1. Error amplification in multi-step workflows

Agentic AI systems “hold immense potential, but inaccuracies or biases in their models can lead to error amplification”. A single error in a multi-step process can propagate and magnify throughout the decision-making chain, especially when real-world complexities, such as cultural or situational nuances, are not considered.

This can result in:



Ethical concerns: Amplified bias and unfair treatment, particularly of vulnerable groups

Erosion of trust: Repeated inaccuracies undermine confidence in AI systems

Operational failures: Errors in critical applications such as healthcare, financial markets, or autonomous operations

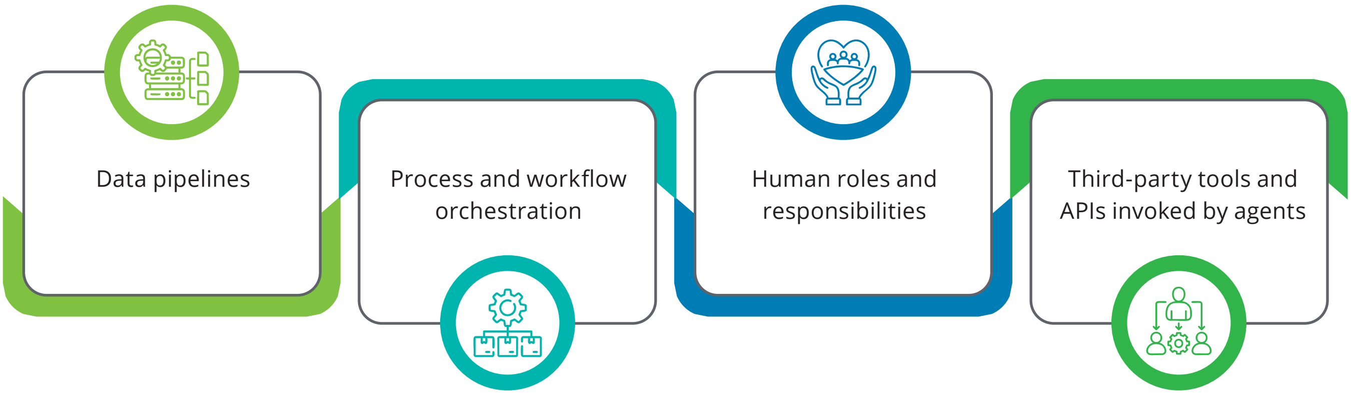
Legal and compliance risks: Regulatory penalties and litigation where AI-driven decisions are not defensible

These risk categories directly mirror the types of harms identified in NIST AI RMF and OECD AI Recommendations, highlighting robustness, accountability and respect for human rights.

2. Socio-technical complexity

Agentic AI operates within complex socio-technical environments where people, policies, processes and technologies are deeply intertwined. Design methodologies must recognise that autonomy, adaptability and interaction need to be complemented with explicit design principles that ensure trust and limit unexpected behaviour.

This demands a System-of-Systems (SoS) view. Guardrails should apply to the model as well as to:



Data pipelines

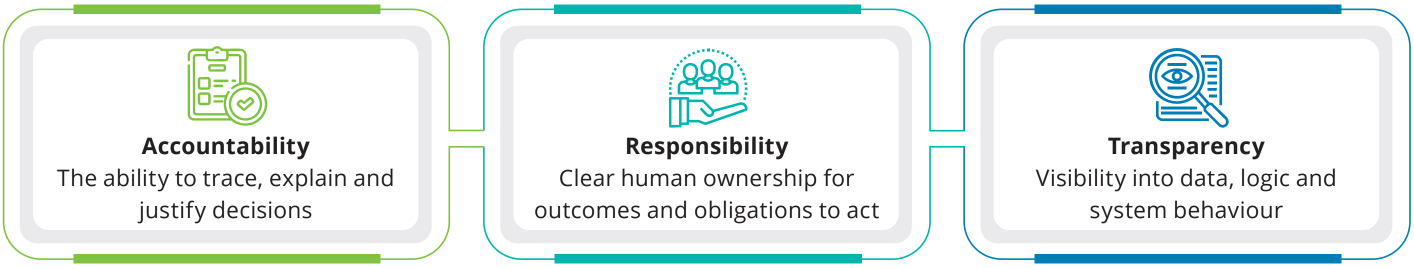
Process and workflow orchestration

Human roles and responsibilities

Third-party tools and APIs invoked by agents

Principles for responsible agentic AI by design

Enterprises can embed responsibility into agentic AI from day one by following these three interconnected principles, consistent with global trustworthy AI standards:



Accountability
The ability to trace, explain and justify decisions

Responsibility
Clear human ownership for outcomes and obligations to act

Transparency
Visibility into data, logic and system behaviour

These principles underpin a socio-technical framework that blends software, governance and regulation across each stage of the lifecycle.

1. Accountability: Traceable agents and explainable outcomes

To be accountable, an agentic AI system must be able to "give account" of its actions and decisions, before, during and after events.[10,11]

Key design patterns:

- End-to-end logging and tracing
  - Each agent interaction records: Prompts, context, retrieved data, decisions and downstream tool calls.
- Explainable reasoning chains
  - Ability to reconstruct why a specific recommendation, escalation or action occurred.
- Post-incident analysis
  - "Black box" style logs enable investigation when things go wrong and contribute to continuous improvement.

Documentation, interpretability and traceability are highlighted by the NIST AI RMF as central to mitigating AI risk.

2. Responsibility: Humans remain on the hook

Responsibility refers to the role of people in relation to AI systems and the need for mechanisms that connect system decisions to human stakeholders.

Guardrails include:

- Role clarity and RACI
  - Defined responsibilities for product owners, AI leads, risk/compliance, operations and business users.
- Human-in-the-loop/on-the-loop
  - High-impact, high-risk decisions require mandatory human oversight and sign-off.
- Escalation and override mechanisms
  - Users should be able to pause or override agent actions without friction, with support provided rather than penalties.

Responsibility is not about "making machines ethical"; it is about designing the socio-technical system so that people remain empowered to intervene, correct and improve.

3. Transparency: Visibility into data, decisions and design choices

Transparency is the ability to describe, inspect and reproduce the mechanisms by which AI systems make decisions and learn, as well as their data provenance.

For agentic AI, this implies:

- **Transparent data pipelines:** Documented sources, transformations and retention policies for all data in the knowledge and vector layers.
- **Policy-aware agents:** Agents are built with explicit constraints (e.g., jurisdictions, thresholds, PII handling) that can be inspected and tested.
- **Stakeholder visibility:** Regulators, internal auditors and impacted functions can review system design choices and their implications.

Operational guardrails for agentic AI deployment

Translating principles into practice requires concrete guardrails that can be applied across use cases and industries.

1. Lifecycle guardrails

a. Design and build

- Risk-based use case qualification and tiering
- Model selection and fine-tuning with fairness and robustness tests
- Secure design of MCP, A2A and ACP integrations

b. Deploy and integrate

- Controlled release through CI/CD and AgentOps pipelines with ethics/risk gates as well as technical tests
- Sandbox and shadow modes before full autonomy
- Role-based access controls to sensitive capabilities (payments, approvals and customer communications)

c. Monitor and improve

- Continuous evaluation of performance, bias and drift
- Periodic reviews of logs and escalations to refine prompts, policies and safeguards
- Feedback loops from users, risk and regulators for iterative improvement

2. Architectural guardrails

To support multi-agent systems, human-in-the-loop, memory management and hybrid reasoning, responsible architectures should:

- Constrain autonomy by design
  - Agents act within explicit guardrails (e.g., transaction limits, approved playbooks) rather than unbounded improvisation.
- Combine probabilistic and deterministic logic
  - Hybrid reasoning blends LLM creativity with rule engines and knowledge graphs for safety-critical decisions.
- Segment responsibilities across specialised agents
  - For example, extraction agents, vetting agents and orchestration agents, each with narrow, testable behaviours and clear accountability.

3. Organisational guardrails

Finally, responsible agentic deployment is anchored in culture and governance:

- AI governance council/CoE
  - Oversees standards, approves high-risk use cases and ensures alignment with organisational values and regulations.
- Policies and codes of conduct
  - Guidance for developers, product owners and end-users on acceptable and unacceptable uses of agentic AI.
- Change management and workforce readiness
  - Structured programmes to build GenAI/agentic literacy and adoption while managing concerns, expectations and ethical awareness.

Way forward: From pilots to trusted autonomy

Agentic AI is a new operating layer for the enterprise. As agents are entrusted with more decision rights and operational reach, responsible design and deployment become non-negotiable.

Across this series, organisations have been guided through six critical questions, from identifying the right use cases and assessing technology readiness to measuring impact, scaling strategically and enabling human-agent collaboration. This closing chapter brings those threads together with a single imperative: to design agentic AI that is responsible from the start, through architecture, governance and culture, not as an afterthought.

Enterprises that take this approach will avoid pitfalls such as error amplification, bias and regulatory backlash while building a trusted, resilient and differentiated agentic AI capability that serves as a core engine of transformation in the years ahead.

References

1. “The Business Imperative for Agentic AI,” Deloitte India, 2025. [Online]. Available: https://www.deloitte.com/in/en/services/consulting/services/engineering-ai-data/the-business-imperative-for-agentic-ai.html

2. National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. Gaithersburg, MD, USA: U.S. Dept. of Commerce, Jan. 2023. [Online]. Available: https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10

3. Organisation for Economic Co-operation and Development, Recommendation of the Council on Artificial Intelligence, OECD Legal No. 0449. Paris, France: OECD, 2019. [Online]. Available: https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

4. OECD, “OECD AI Principles,” 2019, updated 2024. [Online]. Available: https://oecd.ai/en/ai-principles

5. Unlocking the Right Agentic AI Use Cases, in The Business Imperative for Agentic AI series, Deloitte India, 2025. [Online]. Available: https://theshift.info/wp-content/uploads/2025/09/1758826526458.pdf

6. Evaluating Agentic AI Technology Readiness, in The Business Imperative for Agentic AI series, Deloitte India, 2025. [Online]. Available: https://www.deloitte.com/in/en/services/consulting/services/engineering-ai-data/agentic-ai.html

7. Defining and measuring the success of agentic AI, Deloitte India, 2025 [Online]. Available: https://www.deloitte.com/content/dam/assets-zone1/in/en/docs/services/engineering-ai-data/2025/in-eaid-chp-3-pov-defining-and-measuring-the-success-of-agentic-ai.pdf

8. Beyond the pilot: Building, scaling and sustaining agentic AI initiatives, Deloitte India, 2025 [Online]. Available: https://www.deloitte.com/content/dam/assets-zone1/in/en/docs/services/engineering-ai-data/2025/in-eaid-chapter-4-pov-approach-to-building-agents.pdf

9. Building human–agent synergy: Equipping and empowering employees to thrive alongside AI agents, Deloitte India, 2025 [Online]. Available: in-eaid-chp-5-empower-employees-to-thrive-alongside-ai-agents.pdf

10. V. Dignum, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way.* Cham, Switzerland: Springer, 2019.

11. V. Dignum, “Responsible Artificial Intelligence: Designing AI for human values,” ITU Journal: ICT Discoveries, vol. 1, Special Issue 1, pp. 1–8, Sep. 2017.

Connect with us

Ashvin Vellody
Partner
Deloitte India
ashvinv@deloitte.com

Namratha Rao
Partner
Head of Digital Excellence Center
namrao@deloitte.com

Dr. Jagdish Bhandarkar
Chief Disruption Officer
Deloitte India
jbhandarkar@deloitte.com

Moumita Sarker
Partner
Deloitte India
msarker@deloitte.com

Contributors

Namaratha Rao
Dr. Jagdish Bhandarkar

Acknowledgements

Neha Kumari
Ruchira Thakur
Srishti Deoras
Sagarika Mamik Gupta

Deloitte.

This communication contains general information only, and none of DTTL, its global network of member firms or their related entities is, by means of this communication, rendering professional advice or services. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser.

No representations, warranties or undertakings (express or implied) are given as to the accuracy or completeness of the information in this communication, and none of DTTL, its member firms, related entities, employees or agents shall be liable or responsible for any loss or damage whatsoever arising directly or indirectly in connection with any person relying on this communication.