AI Agents in Action: Foundations for Evaluation and Governance *World Economic Forum & Capgemini (2025)*

Context

AI agents are moving from prototypes to real-world deployment, yet most organisations remain unsure how to evaluate, manage and govern them responsibly. With 82% of executives planning adoption within 3 years, the gap between rapid experimentation and mature oversight is widening. This white paper proposes a four-pillar framework across classification, evaluation, risk assessment and governance. It helps adopters scale human-AI collaboration with proportionate safeguards rather than retrofitting controls after incidents occur.

Key Insights

As agents act with varying autonomy and authority, trust should be built incrementally through sandbox testing controlled pilots before full rollout.

Figure 10 below shows how to define the use before evaluating the agent, from classification and evaluation to risk assessment. Table 2 shows key governance mechanisms in practice.

FIGURE 10 | **Foundations for AI agent evaluation and governance: progressive governance practices**

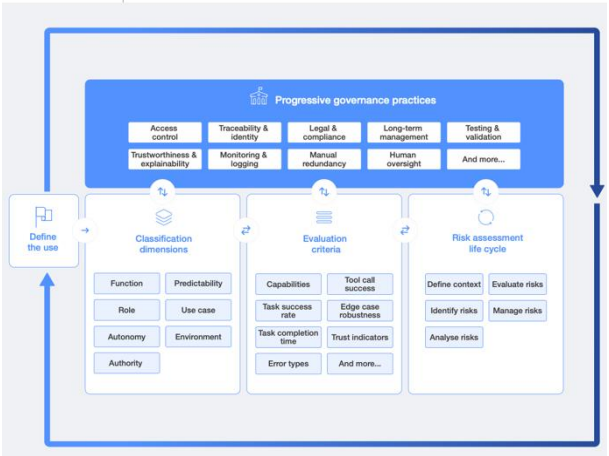

TABLE 2 | **Baseline governance mechanisms for AI agents**

Governance area	Foundational mechanism	Purpose
Access control	Enforce least-privilege access; define task boundaries.	Prevent each agent from accessing unnecessary data, systems, or tools; reduce risk of misuse or accidental harm.
Legal and compliance	Conduct a data protection impact assessment (DPIA); perform privacy and regulation compliance checks, such as General Data Protection Regulation or the California Consumer Privacy Act (CCPA).	Ensure data handling and processing complies with relevant laws and regulations.
Testing and validation	Perform sandbox runs or controlled pilots with non-production data; install input-output filters; perform third-party audits.	Validate expected behaviour, detect errors and prevent untested code from affecting live systems, conduct audits (code, red teaming, etc.).
Monitoring and logging	Implement logging for all agent actions; set up anomaly alerts or dashboards.	Maintain traceability for accountability; enable early detection, incident response and post-incident analysis.
Human oversight	Define and assign oversight models, including HITL/HOTL. Require policy review before deployment and set supervisory triggers for exceptions.	Ensure accountable human control for material decisions, keep behaviour aligned with organizational policies and provide escalation paths when the agent acts unexpectedly.
Traceability and identity	Assign unique agent identifiers; tag outputs to the responsible agent instance.	Attribute actions and outcomes to specific agents; enable forensic review and performance tracking.
Long-term management	Establish protocols for ongoing monitoring, updates and eventual decommissioning.	Ensure continued alignment, performance and relevance throughout the agent's life cycle.
Trustworthiness and explainability	Implement explainability tools; establish trust metrics.	Ensure agent behaviour is interpretable and measurable; build user confidence.
Manual redundancy	Establish manual redundancy procedures to ensure the sustained continuity of critical business use cases.	Preserve data integrity and plan for human resources to take over.

Implications

- How would you classify, evaluate and assess your existing or planned agents (Figure 10)?
- Where does your current agent governance have gaps (Table 2)? How could you address them?
- How will you demonstrate that trust in your agents has been earned rather than assumed?
- What would trigger you to reduce an agent's autonomy or authority in real time, and is that escalation path formalised? Where does accountability sit between your agent providers and your internal deployment teams, and is that split documented?

World Economic Forum. (2025). AI agents in action: Foundations for evaluation and governance [White paper]. In collaboration with Capgemini. https://www.weforum.org/publications/ai-agents-in-action-foundations-for-evaluation-and-governance/