AI Governance for Agentic Systems:
Governance Must Act before AI Starts Acting

From Governing Models to Governing Machines: The New AI Governance Challenge

By Brajesh De, Managing Director at Blue Altair

INTRODUCTION

Most enterprises are building their AI governance for a world where AI generates things. As an example, a Generative AI assistants draft emails, summarize a document or recommend a next step. A human reviews and decides whether to accept and act on it or reject it. But Agentic AI goes beyond that contained transaction. It interprets a goal, reasons through a problem, invokes tools, pulls enterprise data, and increasingly executes the action itself - like processing a refund rather than proposing one, or initiating a payment rather than suggesting it.

This is a significant change in capability of AI from an author to an actor. It moves the central governance question from 'can we trust the model and its output for accuracy, bias, and safety' to 'can we trust what the AI agent is allowed to do, not just its suggestions but what it is authorised to touch, change, and execute?' Agentic AI introduces new dimensions to the AI governance stack that the stack was never designed to answer.

Every CTO, CIO, or enterprise architect scaling Agentic AI must be prepared to align to this shift and close the governance gaps. Governance for agents has to work in both directions: proactively, so an agent cannot go rogue in the first place, and reactively, so incidents are detected, contained and corrected fast. If the gaps are not closed, it can fail their agentic AI projects.

Generic enterprise governance assumes a human actor who can be trained, supervised and held accountable. AI governance extended that to a model whose outputs are probabilistic, so the guardrails focused on accuracy, bias, privacy and explainability.

WHY TRADITIONAL AI GOVERNANCE IS NO LONGER ENOUGH

The traditional AI governance framework already addresses areas such as model risk, privacy, bias, explainability, security, compliance, and human oversight. These remain essential for Agentic AI, but it introduces new governance challenges that traditional frameworks do not adequately address. Agentic AI governance has to govern software that chooses its own sequence of steps, holds standing access to enterprise systems and acts without a human reviewing each transaction. Hence, it adds six things that traditional AI governance never covered viz, **Autonomy, Intent,**

Permission, Delegation, Interaction and Persistence as shown in the figure below.

Hence, Agentic AI needs not just a stricter governance, but a whole new kind of governance, that the traditional AI governance never addressed.

Figure 1: Six dimensions of Agent Authority

THE NEW GOVERNANCE OBJECT: THE AGENT, NOT JUST THE MODEL

An Agentic AI involves an ecosystem consisting of model, prompt, memory, tools, data access, APIs, policies, other agents, and human oversight, stitched together. Hence only governing the underlying LLMs is insufficient.

Governing an agent by governing its model alone is like governing an employee by evaluating only the employee's IQ while ignoring their job description, system access, authority, incentives and actions.

THE SIX DIMENSIONS OF AGENTIC AI GOVERNANCE

Agentic AI governance must organise itself around the agent as a system, not the model as an artefact. It must therefore address the following six dimensions as shown in the figure below:

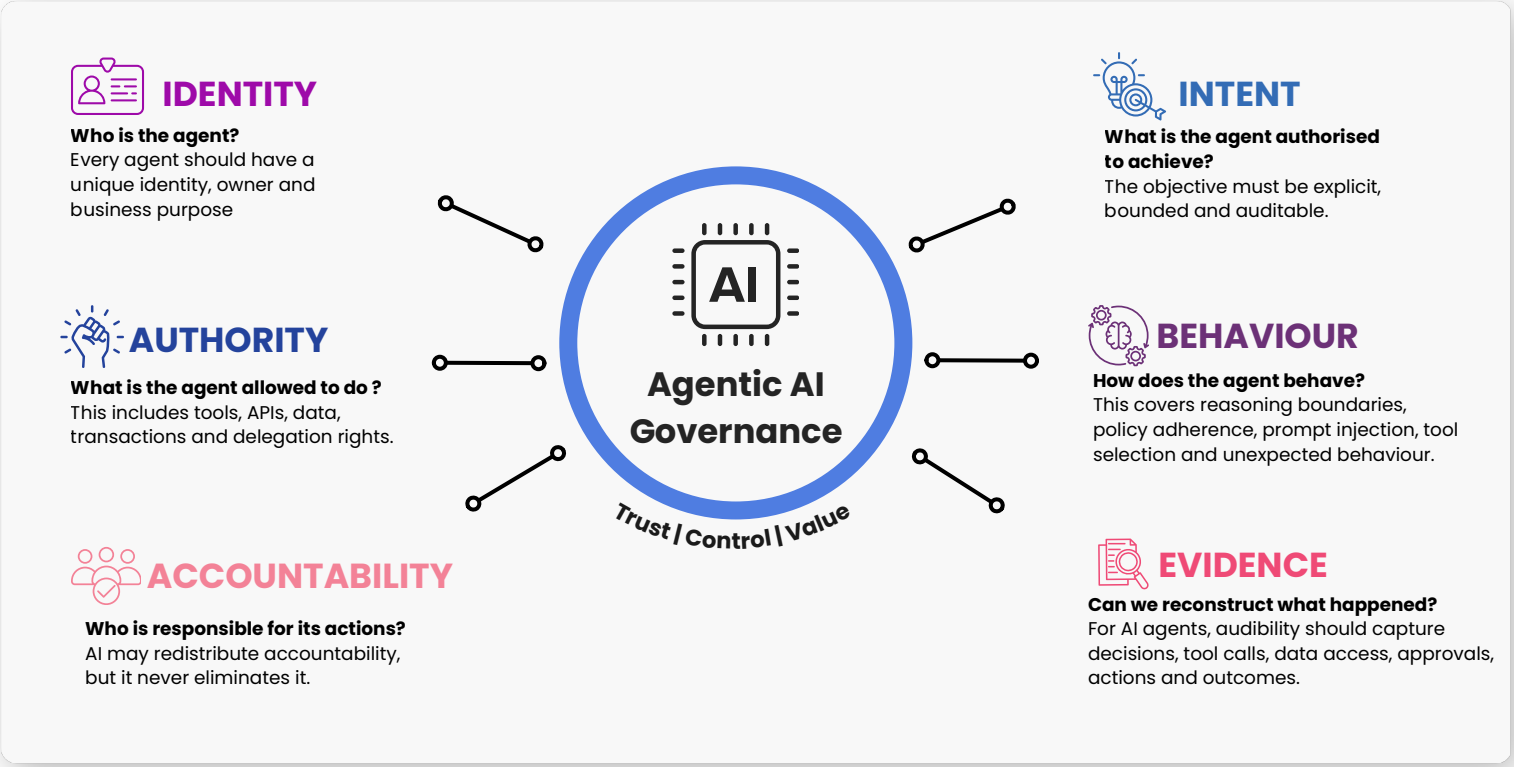


Figure 2: Dimensions of Agentic AI Governance

GOVERNANCE MUST BECOME DYNAMIC

Agentic systems evolve as models, tools, policies and operating environments change. A static annual governance review is no longer adequate. Applying the same governance to every AI agent equally is likely to cause enterprise adoption to fail. Agentic AI governance must therefore adapt to changing conditions and remain proportionate to the risk presented. It should be principle-based rather than rule-bound, so every new tool does not require a policy rewrite. It must also scale from a few showcase deployments to hundreds of agents. Most importantly, it should continuously improve based on observed incidents and near-misses.

Governance should be continuous, but intervention should be risk proportionate. A low-risk internal research agent should not face the same controls as an autonomous agent capable of initiating financial transactions. The governance decision should be based on the impact and the autonomy of the agents. The higher the impact and the greater the autonomy, the stronger the governance must become.

FROM HUMAN-IN-THE-LOOP TO HUMAN-ON-THE-LOOP

Traditional governance often assumes a human approval for every AI recommendation before it is acted upon. However, this can be a bottleneck. As Agentic AI moves forward, humans should monitor AI outcomes and guide AI to take actions. It needs a different form of oversight, suited for systems that need to operate at machine speed and machine scale. This requires organizations to define where humans must approve, where they merely supervise, and where autonomous execution is acceptable.

The key executive governance question now becomes:
Which decisions can an agent make, which decisions can it recommend, and which decisions must remain human decisions?

Governance needs an Agent Permission Architecture

Instead of giving agents broad access, organizations should establish an agent permission architecture with the following capabilities and guardrails:

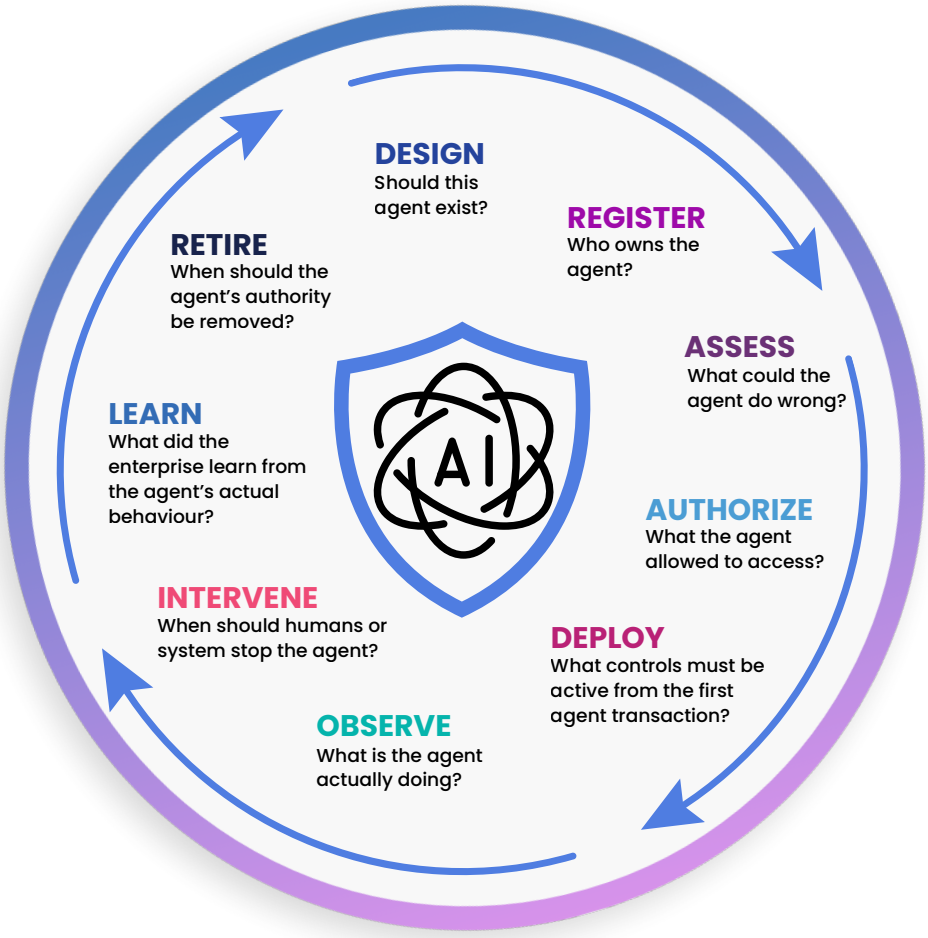
- Enforces least privilege by default, based on the agent's specific roles and tasks
- Provides deliberate authorization for access to specific tools and data
- Caps the transaction limits for any action an agent can take without escalation
- Provides time-bound permissions to carry out specific tasks
- Defines risk-based approval thresholds and actions requiring human oversight.
- Controls agent-to-agent delegation for authority handover
- Enables authorized personnel to suspend an agent, revoke its access or initiate rollback or compensating actions
- Assigns an accountable steward to approve each agent's data and tool access, guided by automated recommendations.
- Continuously monitors permission use and records access, delegation, approvals and actions in an immutable audit trail.

These capabilities need an enforcement point. Most enterprises already have one in their API gateway platform, and the MCP gateway extends it into the control point for agent access to enterprise capabilities. Building the permission architecture on infrastructure that already exists is faster than standing up a parallel control plane for agents.

GOVERNING THE AGENT LIFECYCLE

Governance for AI Agents should span the agent's entire existence. It should begin from the decision to build, through the deployment and continue until the agent is retired. Agentic AI Governance should introduce checkpoints at every stage as shown in the figure below:

AI Governance - Agent Lifecycle



THE BIGGEST BLIND SPOT: AGENT ECOSYSTEM RISK

One agent acting in isolation may be of low risk. But with time enterprises will mostly deploy hundreds of agents that invoke common APIs, access shared data and delegate tasks to each other. The enterprise risk now magnifies manifold. The governance challenge therefore shifts from Agent Risk to Agent Ecosystem Risk, where interactions and cumulative effects can create risks that are invisible when agents are assessed individually.

At that scale, the real question stops being “is this agent safe” and becomes “what happens when this agent interacts with fifty other agents and two hundred enterprise systems.”

MEASURING AGENTIC AI GOVERNANCE

Governance maturity is not about the number of policies or approvals. It is about whether governance actually works.
An Agentic Governance Scorecard should measure visibility, ownership, risk classification, permission controls, human oversight, policy violations, incidents, response time, auditability, and agent retirement.
These metrics should also be connected to business value and trust. Governance should reduce risk without becoming a standalone control function that slows innovation.

I**n short:** Measure whether governance is working, not whether governance exists.

CONCLUSION

Agentic AI moves the governance question from trusting a model's output to trusting an agent's authority to act. That requires treating the agent as a system, not a model; making governance continuous and risk-proportionate rather than static and uniform; deliberately choosing where humans approve versus merely monitor; and building the permission architecture, often through the same API and identity infrastructure enterprises already run, that actually enforces those decisions at scale.

The uncomfortable version of the executive question: if your organisation gave its AI agents the standing access and authority of a new employee tomorrow, would your governance systems actually be ready to manage them? For most enterprises today, the honest answer is no. That is why governance needs to be designed into the agent architecture from day one, with controls that can evolve as autonomy and risk increase.

The time to design Agentic AI governance is before the agents start acting, not after.

About the Author

Brajesh De

Managing Director, APIM and DAD Capabilities at Blue Altair

Brajesh is the Managing Director at Blue Altair. He brings over 25 years of experience across technology, leadership, consulting, architecture, and design, helping large enterprises solve complex technology and business transformation problems. He has led several large-scale initiatives across digital transformation strategy, enterprise architecture consulting, API, Integration, cloud, AI-enabling organizations to scale innovation and business growth.

He has authored two books and holds two published patents in API assessment and data intelligence. Previously, he led API, microservices, and cloud-native capabilities at Accenture, supporting sales, delivery, thought leadership, and reusable delivery assets and accelerators across complex enterprise client engagements globally.