CONCIERTO | SOVEREIGN INTELLIGENCE

Sovereign Perspectives · Verification, Continuity and Trust

Risks, Accountability, and Governance in Agentic Landscapes

A Framework for the Four-Stage Agentic Progression and the Six Risks It Raises

Series Sovereign Perspectives
Category Verification, Continuity & Trust

Format Brief
Published by Concierto



Version 1.0 · September 2026

Copyright © 2026 Concierto. All rights reserved.

Contents

Executive Summary

1. The Progression, Not the Leap
2. The Risk Framework
3. Why the Chain of Responsibility Breaks
4. If This Goes Unattended
5. The Path Forward: A Deterministic Control Layer
6. What Leaders Must Do
7. The Choice

About This Series · Sources

Executive Summary

Agentic AI is not a single leap but a progression, and at each stage the surface of risk widens faster than the governance built to contain it.

The move from AI that answers to AI that acts unfolds in stages: agents first widen their coverage within a function, then begin passing work to one another inside an organization, then transact across enterprises, and finally take consequential actions in the world. Each stage adds capability, and each adds a distinct class of risk that the previous governance did not have to consider. A probabilistic model that produces a plausible, engaging answer is a manageable thing to review; an autonomous agent that takes an action, or sets other agents in motion, breaks the chain of responsibility unless that chain was designed in. This paper offers a framework: the risk classes that emerge across the progression, how a rented and shared architecture compounds each of them, and how a sovereign, deterministic architecture minimizes each in turn.

In Brief

- The agentic future arrives as a progression: coverage expansion, intra-organization agent-to-agent conversation, inter-enterprise agentic conversation, and expansion into real-world actions.
- Each stage introduces a distinct risk class; risk does not add across the stages, it compounds.
- Probabilistic, unattributable output breaks the chain of responsibility the moment an agent acts.
- A rented, shared architecture compounds every risk class; a sovereign, deterministic one bounds each.
- Accountability becomes a property of the system when every output carries its lineage: the policy, the identity, and the decision trace behind it.

When the machine acts, accountability cannot be reconstructed after the fact. It has to be built into the action before it is taken.

Copyright © 2026 Concierto. All rights reserved.

1. The Progression, Not the Leap

Agentic AI is often discussed as a single arrival. It is better understood as four stages, each of which changes the risk picture.

The first stage is coverage expansion: agents take on more of the work within a single function, still largely bounded by that function's systems and data. The second is intra-organization agent-to-agent conversation: agents begin passing work and context to one another to complete end-to-end workflows, so an error in one becomes an input to the next. The third is inter-enterprise agentic conversation: agents transact across organizational boundaries, where no shared standard governs what each party's agent may do or disclose. The fourth is expansion into actions: agents take consequential, sometimes irreversible steps in the world. Governance that was adequate at the first stage is not adequate at the fourth, and most institutions will pass through all four faster than they expect.

Risk in an agentic estate does not add stage by stage. It compounds, because each stage builds on the exposures of the one before it.

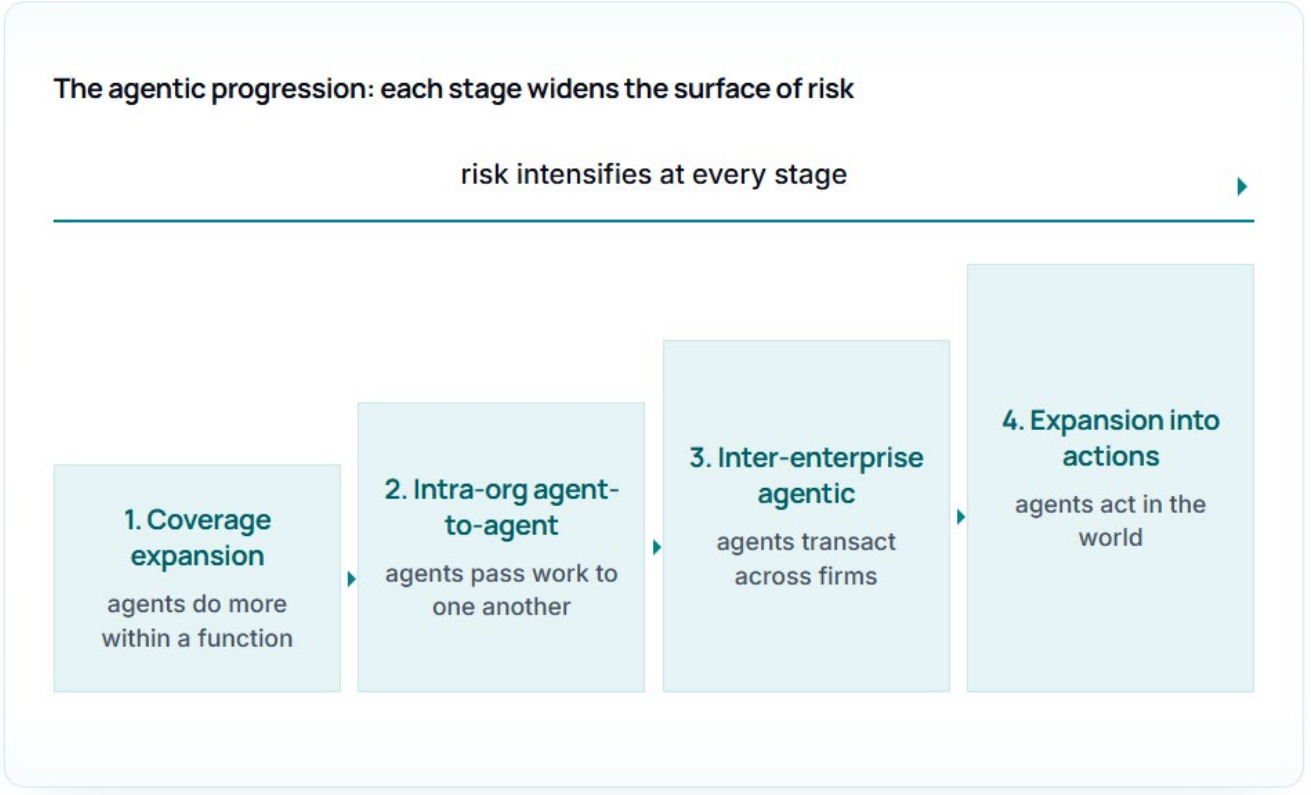


Figure 1. The four-stage agentic progression. Each stage widens the surface of risk. Illustrative.

Copyright © 2026 Concierto. All rights reserved.

2. The Risk Framework



A framework is only useful if it names the risks precisely. Six classes recur across the progression, and each behaves differently under a rented model than under a sovereign one.

The value of naming the classes is that it turns a vague unease about agents into a set of specific exposures a leader can govern, each with a clear contrast between how a rented, shared architecture compounds it and how a sovereign, deterministic architecture minimizes it.

The same six risks run through every agentic estate. What differs is not the risks but the architecture that either compounds or contains them.

Risk class	How the rented, shared model compounds it	How sovereign architecture minimizes it
Data egress and leakage	Connected, shared model; data leaves the perimeter and may train the supplier	No egress by design; sandboxed by topology; nothing leaves the boundary
Unauthorized or unintended action	Probabilistic output with no gate in the action path	Deterministic control layer gates every action before it executes
Cascading agent-to-agent failure	Opaque swarms on shared infrastructure; errors propagate unchecked	Owned orchestration; bounded, logged, interruptible interactions
Attribution and accountability loss	Black-box output, no lineage, no one to hold to account	Every output carries policy, identity, and a decision trace
Inter-enterprise trust and liability	No shared standard; each party's exposure is unknowable	Attested sovereign boundaries; verifiable claims across parties
Emergent misalignment and drift	Scale plus autonomy with no deterministic brake	Rules-based enforcement in the path; human-in-control as anchor

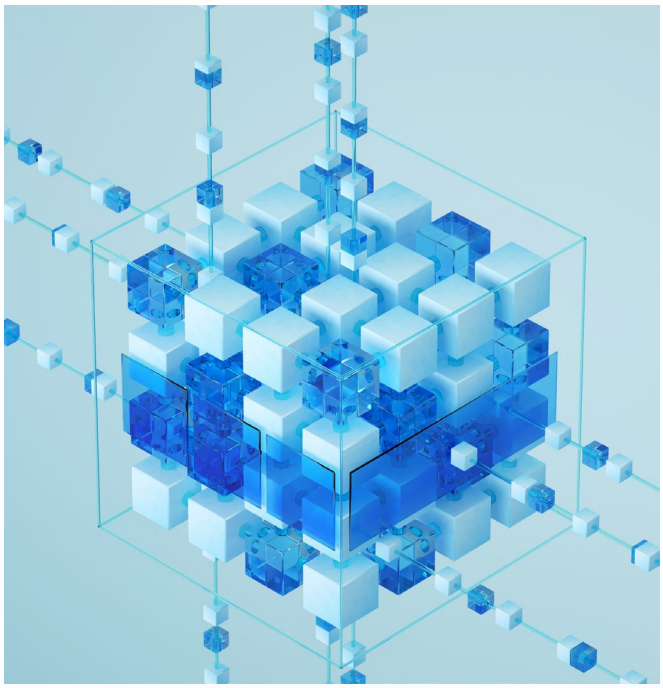
Copyright © 2026 Concierto. All rights reserved.

3. Why the Chain of Responsibility Breaks

The deepest of the six is attribution, because when it fails, none of the others can be governed at all.

A probabilistic model is designed to produce plausible, fluent, engaging output. That is a virtue when a person is reading and deciding, and a liability when an agent is acting, because plausibility is not permission and fluency is not accountability. When an autonomous action goes wrong and the output that caused it cannot be traced to a policy that permitted it, an identity that authorized it, and a decision trace that produced it, there is no one to hold to account and nothing to correct. Responsibility diffuses across the model provider, the application builder, the agent's author, and the operator, each able to disclaim it. Accountability cannot be added back after the incident; it has to be a property of the action when the action is taken.

> **"Systems that take consequential actions must be traceable to the decisions and the people behind them, or they cannot be trusted with those actions at all."**
> OECD, Recommendation of the Council on Artificial Intelligence · 2024. [2]



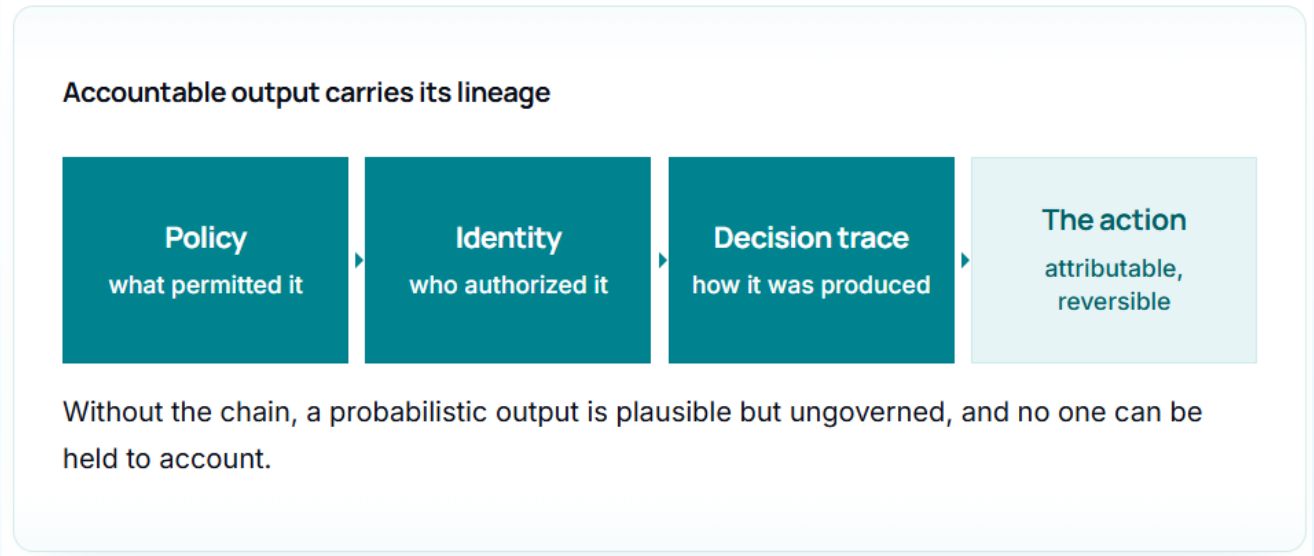


Figure 2. Accountable output carries its lineage. Without the chain, a plausible output is ungoverned.

Copyright © 2026 Concierto. All rights reserved.

4. If This Goes Unattended

Run the progression forward on a rented, shared foundation and the six risks do not stay contained to a function. They compound across the estate.

Left unattended, coverage expansion becomes an unmonitored dependency; intra-organization agent-to-agent conversation becomes an untraceable web where one agent's error is another's input; inter-enterprise agentic conversation becomes an exchange of actions between parties who cannot verify each other's controls; and expansion into actions becomes irreversible steps taken by systems no one fully governs. As liability regimes form around AI decisions, an institution that cannot attribute an action cannot defend it, and an institution whose acting intelligence is rented cannot even promise a regulator that the controls will not change beneath it.

Unattended, an agentic estate becomes a set of actions no one can trace, taken by systems no one fully controls, under a liability regime that will nonetheless assign the blame.

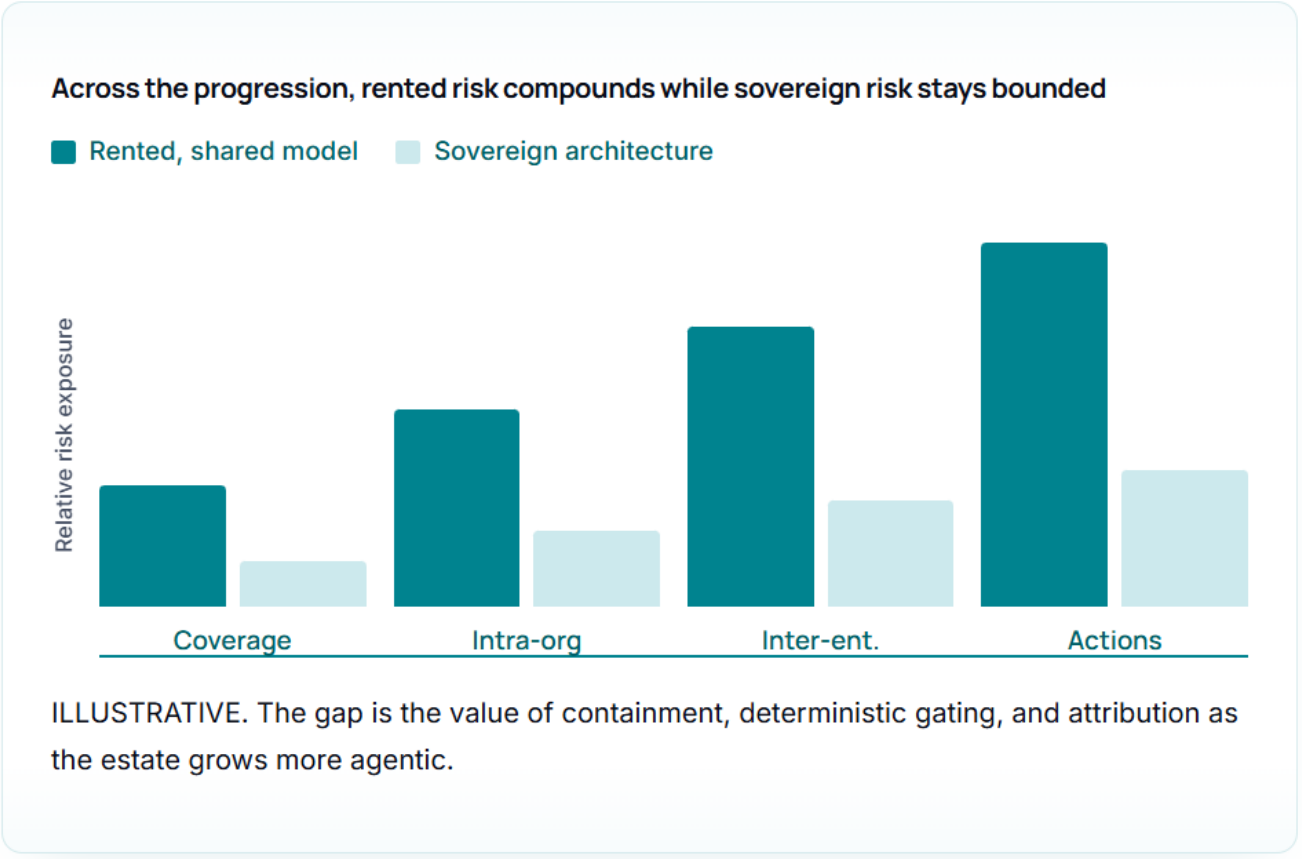


Figure 3. Across the progression, rented risk compounds while sovereign risk stays bounded. Illustrative.

Copyright © 2026 Concierto. All rights reserved.

5. The Path Forward: A Deterministic Control Layer

Every one of the six risk classes is bounded by the same architectural answer: a deterministic control layer over every output, inside a sovereign boundary.

The defense layer sits between the probabilistic model and the world it would act upon. It is deterministic, so whether an action is permitted is decided by explicit, auditable rules rather than by the model itself; no probabilistic system sits in the enforcement path. It is attributing, so every output leaves with its lineage attached. It is contained, so the agent operates within a sovereign boundary it cannot exceed and from which data cannot leave. And it is anchored by a person, so the authority to permit, refuse, and reverse rests with a human and is evidenced in the record. Read against the framework, this single architecture is what turns each risk class from compounding to bounded, which is why the contrast in the table is not a list of features but one design decision applied six times.

Six risks, one architectural answer: a deterministic, attributing, contained control layer turns every one of them from compounding to bounded.



6. What Leaders Must Do

The progression is already underway in most estates. The framework is useful only if it is acted on before the later stages arrive.

Build the accountability chain before scaling agents, not after an incident forces it. Require that every autonomous output carry its lineage; place a deterministic control layer in the action path; contain agents within sovereign boundaries; and hold human authority over what agents may do, evidenced in the trace. Map the estate against the six risk classes and the four stages, and govern to the stage the estate is heading toward, not the one it is in.

Copyright © 2026 Concierto. All rights reserved.

ACTION	WHAT IT BOUNDS
Require output lineage	Attribution loss: every action ties to policy, identity, and trace
Deterministic gating in the action path	Unauthorized action: permission is provable, not probabilistic
Contain within sovereign boundaries	Data egress and cascading failure: the blast radius is bounded
Attest boundaries across parties	Inter-enterprise liability: claims are verifiable, not assumed
Anchor human authority	Emergent drift: a person can permit, refuse, and reverse

Govern the stage the estate is heading toward, not the one it is in. By the time the risk is visible, the action has already been taken.



Copyright © 2026 Concierto. All rights reserved.

7. The Choice

The agentic future is not a question of if but of how governed, and the framework makes the choice concrete.

An institution can let the progression run on a rented, shared foundation and meet each new risk class as an incident, or it can adopt the deterministic, sovereign architecture that bounds all six from the start. The framework is the same either way; only the architecture decides whether the risks compound or stay contained. The accountable output is not a constraint on the agentic future. It is the precondition for being allowed to have one.

The agentic estate that can trace and govern every action will be trusted to take them. The one that cannot will be stopped, by incident or by regulator.

About This Series

Sovereign Perspectives is a research program on the ownership and governance of intelligence, published by Concierto. Its papers examine the strategic, economic, and institutional questions that artificial intelligence raises for governments and other accountable institutions, and they are written for the leaders making those decisions. The program's aim is to advance the public argument, not to promote a product.

Sources

- Stanford Institute for Human-Centered AI, Artificial Intelligence Index Report 2025. hai.stanford.edu.
- OECD, Recommendation of the Council on Artificial Intelligence, 2024. oecd.ai.
- OECD.AI Policy Observatory, AI incidents and governance instruments, 2024–2026. oecd.ai.

reach@concierto.cloud

www.concierto.cloud

www.trianz.com

Copyright © 2026 Concierto. All rights reserved.