Introduction

The failure mode nobody is watching for

A regional insurance carrier in the Netherlands deploys an agentic AI system to adjudicate claims for one product line. The system was built by a reputable vendor, passed a six-month pilot, and has been in production for a year. The carrier did something most deployers do not: before go-live, it interviewed its senior adjusters and wrote their rules into the agent's instructions, including the one written three years earlier, after a fraud ring had cost it seven million euros, which says that any fourth claim in a sixty-day window from one claimant goes to the fraud team, regardless of amount or score.

A claim arrives. The agent retrieves prior incidents from the customer's record. It verifies that the policy is active and the premium current. It consults the fraud model, which scores the claim as low-risk. It confirms the amount is within the auto-approval threshold. It counts the claimant's claims in the last sixty days, as instructed, and finds two. Every check passes. Every rule is applied. Every log line is green. The agent approves the claim and the payout is scheduled.

Six weeks later, the fraud team opens an investigation. The claim was the fourth from the same household in six months: two in the wife's name and two in the husband's, on two policies at one address, all in the same incident category, each just below the threshold, with sixty-three days between the second and the third. The rule said sixty days and one claimant. It did not fire. It had been followed perfectly, and what it was for had been missed completely.

The second policy was in the same system, at the same address, one query away. The agent never looked, because nothing had told it that the household was something to consider. An adjuster with twenty years in the job would have looked, not because she would have counted differently, but because what she knew was never "sixty days, one claimant." What she knew was what a ring looks like. The sixty-day rule was the carrier's attempt to write that knowledge down. It was a reasonable attempt. It was a compression. The agent had the compression, not the knowledge.

The agent did its job. It followed its rules. The decision was wrong.

This book is about that specific kind of wrongness. I call it *decision-basis failure*: the workflow runs correctly, every operational check passes, and the decision is

still wrong, because nothing verified whether what should have mattered was considered. It is not a bug in the agent. It is the absence of institutional judgment from the chain, and it passes silently through every layer built to catch the other kind of failure. The carrier in this story did not have the architecture this book describes. Chapter 1 returns to the same kind of decision at a carrier that did.

The debt that is accumulating

One such incident is survivable. The carrier absorbs the loss, patches the workflow, moves on. What is not survivable is what accumulates underneath the incidents, invisibly, in every organization deploying agents. I call it *governance debt*, by direct analogy to technical debt. I introduced the term for enterprise AI in March 2026 (Ledesma, 2026); the three questions below are that piece's definition, and Chapter 9 sharpens it into one of three coupled debts. The term has since entered circulation on its own: a CIO.com essay of August 2026 observes that governance debt, unlike technical debt, is "neither patient nor pausable," because agents run in loops without human gates (Siraj & Snodgrass, 2026). That is the right observation, and the parallel uses are welcome; this book adds the mechanism, autonomy outrunning the reference that should govern it, and the two companion debts.

It builds quietly. One team deploys an automation agent. Another adds an AI workflow to a process. A vendor switches on agent capabilities inside a platform the organization already uses. Each deployment was someone's reasonable idea, approved by someone who trusted the team. Over time, three questions that should be easy to answer about any operational system become hard.

Which agents are actually running? Many organizations cannot say. Agents sit inside tools, workflows, copilots and internal applications, often without central registration.

What authority was delegated to them? The boundaries exist somewhere in prompts, code or configuration. There is no governance record of what humans approved.

Who is accountable for their decisions? The agent acted. Under whose authority, with what constraints, nobody can say with confidence.

An organization that cannot answer these has built an operational layer of autonomous systems without the governance infrastructure to manage it. Nothing has visibly failed. The debt comes due the moment something does, or the moment a regulator asks. It is invisible on the balance sheet because there is no balance sheet for governance exposure.

Organizations that take the three questions seriously usually answer them by naming tools they already have: the harness, observability, scope definitions, skill specifications, MCP permissions. Each is real and each solves a slice of the problem. None produces the governance record the questions require, because each describes the operational stack — what the agent can touch, what it did, how its loop ran — and not the judgment that was supposed to shape what it did, or who signed for it. Chapter 3 goes through these tools one by one; Chapter 29 returns to the harness, which carries more of the misplaced confidence than the other four together.

Governance debt is not the only debt agentic AI is running up. Technical debt accumulates faster when agents write the code. Environmental debt accumulates in the energy and water the compute consumes. Social debt accumulates in the populations left behind in the safe use of these systems. Cognitive debt accumulates in the people who hand their judgment to the systems. Institutional memory debt accumulates when an organization can no longer reconstruct why its agents decided what they did. This book is about governance debt, and the architecture it develops pays down that one and, through the Governance Ledger, institutional memory debt. It does not pay down the others, and it should not be built in ways that make them worse. The closing chapter returns to the four it leaves untouched.

The question that decides whether you can trust the outcome

For two decades the industry has grown steadily better at a particular set of questions. Who may act? Authentication. On what? Permissions and policies. With which tools? Orchestration. What happened? Observability. How well did it perform? Evaluations. Together these form the *execution stack*: serious engineering that answers *how*, *when* and *by whom* an action was taken.

None of it answers the question that decides whether you can trust the outcome: **was this decision sufficiently supported to act autonomously?** That is a question about *basis*, not execution. Did the agent weigh what a competent decision-maker inside this specific organization would have weighed? Did it consider the pattern, not just the instance? In the insurance case every execution-stack check was green. The question the stack did not ask was whether the agent had considered the one piece of institutional judgment that mattered, because that question lives at a different layer. The stack sees the workflow: what the agent did, and in what order. It does not see the **decision trajectory**: how the decision developed toward its outcome, and whether that path included what the organization’s judgment requires.

Why this matters now

Decision-basis failure is as old as autonomous decision systems. A credit-scoring model in 2005 could fail this way. What has changed is the arithmetic of exposure around it. Exposure has grown along four axes: *volume*, because agents produce decisions at machine throughput rather than recommendations at human pace; *magnitude*, because an agent moves money, signs contracts and deploys code where a scoring model returned a label; *surface*, because the classes of decision now delegated — claims, releases, procurement, hiring triage, incident response, reconciliation — have multiplied, each with its own topology of institutional judgment; and *propagation*, because agents call agents, and a silent failure early in a chain is processed as truth by every link after it. Detection has shrunk along three: the *intervention window* collapsed from days to milliseconds; the *legibility* of a single wrong decision vanished inside a stream of operationally identical right ones; and the *pattern* stays invisible because each surfaced failure is patched as its own incident.

That is the pincer, and Chapters 2 and 5 work out its arithmetic. The difference in one line: a scoring model was a *recommender*, and a human stood between its output and the world; an agent is an *actor*, and nothing stands there.

Iansiti and Lakhani (2020) described the firm whose operating core is an "AI factory": data pipelines, models and experimentation platforms that remove humans from the critical path of operations, so that scale, scope and learning no longer run into the limits of a human organization. Their account is right about the operating model and silent about the decisions it makes. An AI factory produces decisions the way a plant produces parts, and the question this book asks is the one the factory does not answer: whose judgment are those decisions held to, and who answers when one of them is wrong.

Management and governance, one machinery

One claim runs through this book, and it is more consequential than it first appears. **In agentic organizations, management and governance stop being two layers running at two speeds. They become two authorities acting through one machinery.**

The classical era separated them deliberately. Management ran operations; governance oversaw management, at a lag: the board quarterly, the regulator annually, the audit afterwards. The lag was where oversight did its work, and it presupposed that operations ran at human speed. Agentic organizations have no lag. Oversight that is quarterly is not oversight; it is history. For oversight to mean anything it must be simultaneous with operation, in the same architectural pieces. The Steward Agent that enforces authority *is* the system that evaluates the decision as it is proposed. The Blueprint that specifies

authority *is* the operational specification. The Ledger that creates the audit record *is* the operational log.

What does not collapse is who may change that machinery. Operators run it. Signers authorize it. Boards approve the risk appetite it encodes. The Steward Agents answer to the signers, not to the team whose agents they supervise. The machinery is shared; the authorities are not. Boards, auditors and regulators keep their legal mandates, which differ by jurisdiction and are not the subject of this book. That distinction is what keeps the architecture from becoming management grading its own homework at machine speed.

Three consequences follow for the people implicated. Boards do not disappear; they move from reviewing last quarter to signing and maintaining the Blueprint, with the responsibility that goes with signing. Executives do not disappear; their work shifts from direct decision-making to designing and running the architecture that decides, and adjudicating what it escalates. Compliance does not disappear; it moves upstream, from auditing after the fact to writing the artifacts that make behavior governable in the first place. The reader who finds the governance vocabulary ahead heavy should hold this in mind: it is only heavy if you still read governance and management as two layers.

What this book argues

The argument has seven moves.

First, classical management rested on two assumptions that quietly stopped being true when agents arrived: that decisions can be reviewed before they execute, and that governance has the shape of a pyramid. Part II traces how. Second, the right response is not to patch classical management but to rebuild it on a different unit: the decision itself, with measurable properties — blast radius, basis quality, propagation depth, reversibility, required scrutiny — that Simon in 1947 could not instrument and that agentic environments force into existence. Part III develops this. Third, taking the decision as the unit moves the organization's shape from pyramid to network: a decision graph signed in advance, node by node, by the specific person responsible, against which every agent decision is checked before it executes. Fourth, the pattern is not new; Toyota did it to quality with *jidoka* and DevOps did it to testing with shift-left, and Part IV reads both precedents, and where they stop short. Fifth, the architecture that makes this work has three components — direction, real-time verification and memory — and seven graded interventions rather than two. Part V specifies it. Sixth, the architecture has a generative face as well as a defensive one: configured for dialogue and plural perspective, stewards make the decisions that pass better, not only stop the ones that should not. It also has limits, which Chapter 24 names before the book asks anyone to build it.

Seventh, Part VI is practical: four roles becoming scarce, a playbook, a transition map. Part VII places the whole thing beside cybersecurity, the other defense that draws on the same human attention.

One idea runs through the seven, and it is worth stating before the detail. The reference the organization keeps is live: its owners revise it as incidents, policies and the business change, so that what counts as a well-supported decision is always the organization's current judgment rather than a snapshot. Each part of it carries the logic a decision must satisfy, the scrutiny it gets, and the intervention its owner chose for a deviation. The conditions under which an agent exercises its authority, its scope, its speed, its thresholds, are adjusted from live performance against that reference, so that an agent earns autonomy by staying close to it and loses autonomy by drifting, before anything fails. The people who own the decisions do not review them. They stay aware of decision quality against their own reference point, and they act in two ways that never touch the agent's code: by intervening in a decision, and by changing the conditions under which the agent operates. The organization governs its agents by conditions, not by instructions. The instructions belong to the team that built the agent; the conditions belong to the organization, and it can change them at the speed its judgment changes. The human leaves the loop and enters the control room.

The book closes where the argument began, in Athens, on a hill beside the Parthenon called the Pnyx, where citizens claimed *isegoría*: distributed voice over the decisions that affected them. The pyramidal era interrupted that form. The nodal firm restores it, with technology that lets it run at machine speed.

Who this book is for

Enterprise executives — CIOs, risk officers, general counsel, heads of AI and operations — whose organizations will deploy agents in the next three years, which is nearly all of them. Researchers in management, computer science and information systems, for whom this book names a failure mode, proposes a reconstruction and defends an architecture, and submits all three for critique; the academic development is my doctoral work, and this is where I invite the field in first. Policymakers drafting oversight requirements, who need a vocabulary precise enough to regulate against. Board members and investors, whose duties and capital reach decisions they cannot individually see; two interludes after Part V are written for them. And security leaders, whose discipline this book sits beside rather than inside: cybersecurity defends against adversarial attack, this architecture against decision-basis failure, and neither substitutes for the other.

A note on method

I write as a practitioner. I run MeaningStack, in Maastricht, where we build Steward, infrastructure that implements this architecture in production for enterprise clients. Patents are pending on the architecture itself. Where I can describe something from observation, I do; the sector cases that recur across the book are drawn from deployments and anonymized. Where I have figures of my own, the body states them qualitatively, because the argument does not depend on their size; the notes at the back give the figure observed, what was counted, and how far it should be trusted. Where I rely on the literature, I cite it. This posture is unusual for a management book. It is also the only honest way to write about something being built faster than research can catalog it. The alternative is a book that is complete and five years out of date.

A word on what kind of claim this is. The book begins with a recurring pattern: execution can be correct while the decision basis is deficient. Naming and describing a pattern is theoretical work, and a theory does not become invalid because a large comparative dataset does not yet exist. But it must be clear about what it observes, what it infers and what it predicts, and its logic must not rest on unverified percentages. Mine does not. The architecture is offered as a proposal, open to rival explanations, counterexamples and independent implementations.

Chapters are short and built around one argument or one case each. Read the book in order the first time; return to chapters alone the second. Terms are set in bold where first introduced and collected in the glossary.

We begin where every decision-basis failure begins: the workflow runs, every check passes, and the decision is wrong.

The full pre-release edition of *The Decision* and every chapter are available from MeaningStack. Contact: luciana@meaningstack.com