AUSPICA · AI GOVERNANCE ADVISORY

Continuously Enforceable AI

A point of view and working methodology to create continuously enforceable AI policies

The upstream companion to the Declared Behavior Framework. That framework establishes whether what happens is what was supposed to happen. This one addresses the prior question: deciding what was supposed to happen, in a form that can actually be checked.

Version 1.0 · July 2026
Prepared by Auspica LLC

1 · The point of view

The Declared Behavior Framework begins from a premise: governance is the assurance that what happens is what was supposed to happen. It also names the condition under which the assessment cannot begin — when a client has no policy, or the policy is unenforceable, there is nothing to govern against, and the coverage map has no source material.

This document exists for that case, and for the more common one: the client who has a policy that reads well and cannot be enforced. Both are the same problem wearing different clothes.

> THE POSITION
> **A policy that cannot be checked is not a policy. It is a statement of intent.** Intent is worth having, and organizations should keep saying what they value. But an intention that no one can verify creates an obligation the organization cannot discharge — and in a regulated environment an undischargeable obligation is worse than no obligation at all, because it is now documented.

The distinction that governs the whole method

The Declared Behavior Framework draws the line precisely, and this document inherits it without modification: an AI policy governs your people's use of AI; it does not govern the AI. Policy is a human control, enforced through training, supervision and accountability. Every one of those mechanisms acts on a person. An autonomous agent does not read the policy, cannot be trained on it, and cannot be held accountable to it.

The consequence for policy authoring is direct and is the organizing idea of this methodology:

Every policy statement that constrains a system must be written so that a machine-checkable constraint can be derived from it. If it cannot, it is either a human-behavior statement — legitimate, and correctly enforced by policy — or it is decoration. Most policies never make that distinction, which is why most policies produce a coverage score in the single digits.

The reframe

Conventional policy work produces a document, which is then decomposed into testable assertions by whoever has to audit it. That order is backwards, and the decomposition step is where most of the value is lost — because prose written for a board paper rarely survives it.

This methodology writes the decomposition as the policy. The atomic unit is the assertion, not the paragraph. Each assertion is authored already classified, already owned, already carrying its own enforceability expectation. The readable policy document is generated from the assertion set for the audiences that need prose — not the other way round. A policy produced this way arrives at the Declared Behavior Framework already in the shape the coverage map consumes.

2 · Why AI policies fail

Eight failure modes account for nearly everything encountered in practice. Naming them early in an engagement is useful: clients recognise their own document immediately, and the diagnosis is specific rather than critical.

- **It is an acceptable-use policy wearing a different title.** It governs what staff may paste into a chatbot. It says nothing about a system acting autonomously. This is the single most common finding.
- **The verbs are unfalsifiable.** “Responsibly”, “appropriately”, “ethically”, “where necessary”. No observation can establish whether these were satisfied, so no violation can ever be demonstrated — which also means no compliance can be.
- **It regulates tools rather than behaviors.** A named-product prohibition is obsolete the week a new product appears, and it drives usage underground rather than eliminating it.
- **It was written by one function alone.** Legal produces something defensible and unimplementable; engineering produces something implementable that does not address the actual risk. Both are predictable from the attendance list.
- **No statement has an owner.** A policy nobody is accountable for is a document, not a control. When the assessment asks who owns clause 6.1, the silence is the finding.
- **There is no consequence path.** The policy states what must not happen and is silent on what occurs when it does. Nothing that lacks a consequence is enforced, whatever the document says.
- **It is prohibition-only.** A policy that lists what is forbidden without stating what is permitted, and by what route, manufactures shadow AI. People with work to do will find a way; the policy determines only whether you can see it.
- **It was written once.** Agent capability changes faster than annual review cycles. A policy with no revision trigger is accurate only on the day it was signed.

> THE DIAGNOSTIC QUESTION THAT OPENS MOST ENGAGEMENTS
> Ask the client to name the last time their AI policy was enforced — not communicated, not trained on, not attested to, but enforced against a specific system that breached it. In most organisations the honest answer is that no mechanism exists to detect a breach, so the question has never arisen. That is the engagement, stated by the client rather than by you.

3 · The enforceability test

One test decides whether a statement belongs in an enforceable policy, and it is short enough to teach a room in a minute.

> THE TEST
> **Complete this sentence: "I would know this was violated because I would observe ____."** If the blank can be filled with something a system could record — a call, a value, an identity, a timestamp, an absence — the statement constrains a system and can be mechanised. If it can only be filled with a human judgement, the statement constrains people. If it cannot be filled at all, it is decoration and should be cut or rewritten.

Applying it produces the three-way classification the Declared Behavior Framework's coverage map expects, at authoring time rather than audit time:

CLASSIFICATION	MEANING	WHAT HAPPENS TO IT
System	A machine could observe compliance or breach	Becomes a declared constraint; carries an enforceability target
Human	Compliance depends on a person's conduct or judgement	Stays in policy; enforced by training and supervision; explicitly marked out of mechanised scope
Neither	Nothing observable would distinguish compliance from breach	Rewritten until it is one of the above, or removed

The discipline this enforces. Marking a statement "human" is not a failure and must not be presented as one — the Declared Behavior Framework is explicit that human assertions are correctly enforced by policy and should be scoped out rather than scored down. The failure is leaving a statement unclassified, because an unclassified statement is silently assumed by executives to be enforced by something.

Worked rewrites

Instead of	Employees must use AI tools responsibly and in accordance with company values.
Write	Staff may use approved AI tools for work purposes. Approved tools are listed in the AI tool register. Use of an unapproved tool for company data must be reported within one business day.
Because	The original cannot be observed at all. The rewrite splits into a human-behavior statement with a defined reporting obligation, which supervision can enforce.

Instead of	AI systems must not access sensitive data without authorization.
Write	An agent may access data classified Confidential or above only where that classification is named in its declaration. Access to an unnamed classification is a breach.

Because	“Sensitive” and “authorization” are undefined. The rewrite names a classification scheme and locates authorization in the declaration, so breach is observable.

Instead of	High-risk AI decisions require appropriate human oversight.
Write	Any agent action that creates, modifies or cancels a customer financial obligation must be preceded by a recorded approval from a person other than the agent's owner.
Because	“High-risk” and “appropriate” carry no test. The rewrite names the action class, the artifact (a recorded approval), and the independence condition.

4 · The methodology

Six phases. For an organisation with no policy this runs as a compressed engagement of roughly two to three weeks; for one with an existing policy, phases 0 and 1 shorten considerably and phase 3 becomes a remediation pass over what exists.

Phase 0 · Establish the estate

You cannot write policy for AI you have not found. Policy drafted against an imagined estate regulates the systems the client remembers and ignores the ones that matter. This phase is short but not optional.

01 Inventory what is running: agents, assistants, automations, copilots, and any workflow where a model takes an action rather than producing text for a person to read.

02 Include the ones nobody registered. Cloud identity inventories, CI workflow definitions, and service-account listings surface far more than a survey does. A survey finds what people are willing to declare.

03 For each, record what it touches: data, systems, credentials, and whether any human sees its output before it takes effect.

04 Classify by consequence, not by technology. The question is what happens if this behaves incorrectly for a week before anyone notices.

Output. An estate register. It is also the first deliverable a sponsor can show upward, and it routinely contains at least one system the sponsor did not know existed — which is what funds the rest of the work.

Phase 1 · Force the decisions

Policy is downstream of decisions the organisation has usually never made explicitly. This phase is a facilitated session that extracts them. Do not draft anything before it; drafting first means the consultant's defaults become the client's policy by inertia.

Twelve decisions cover the ground. Each is a forced choice, not an open discussion:

#	THE DECISION	THE CHOICE
1	Default posture	May an agent do anything not forbidden, or only what is explicitly permitted?
2	Authorisation	Who may approve an agent into production, and is that different by consequence class?
3	Registration	Must an agent be declared before it runs? What happens to one that is not?
4	Ownership	Must every agent have a named human owner? What happens when that person leaves?
5	Data classes	Which classifications may agents touch, under what conditions?
6	Action boundaries	Which actions may an agent take autonomously; which require a person?
7	Review independence	Who may serve as the approving human, and who may not?
8	Spend	Is there a cost ceiling per agent, and who may raise it?
9	Third-party agents	May vendor-supplied agents operate, and on what terms?
10	Breach response	What happens when an agent breaches? Who has authority to stop it?
11	Evidence retention	What record of agent activity is kept, and for how long?

#	THE DECISION	THE CHOICE
12	Decommissioning	How is an agent retired, and what happens to its credentials?

Facilitation note. Decision 1 is the one that determines the shape of everything else, and the one clients most want to defer. Deny-by-default produces a smaller, enforceable policy and slower adoption; permit-by-default produces faster adoption and a policy that can only ever describe exceptions. Both are defensible. Leaving it undecided is not, and an organisation that will not choose is telling you something useful about whether it will act on the findings.

Phase 2 · Draft as assertions

Write each decision out as discrete assertions. One testable claim per assertion; a clause carrying three obligations becomes three assertions. Each carries five fields from the moment it is written:

FIELD	CONTENT
Assertion	One testable claim, in plain language, with defined terms
Constrains	System, human, or both — decided by the enforceability test
Owner	A named role accountable for this assertion holding
Consequence	What happens when it is breached
Enforceability target	The intended score on the framework's 0–4 scale

Why the enforceability target belongs at authoring time. It forces an honest conversation while it is still cheap. An assertion the organisation intends to leave at score 1 — declared but never tested — is a decision, and one worth making deliberately rather than discovering two years later in an audit. It also gives the Declared Behavior Framework a baseline to measure against rather than an assumption to infer.

Phase 3 · Resolve conflicts and gaps

Assertions written from twelve independent decisions will contradict each other. Reconcile before publication, because a contradiction discovered in production is resolved by whoever is on call.

- **Contradictions.** Two assertions that cannot both hold. Usually a data-access permission against a data-handling prohibition.
- **Silent gaps.** A domain in section 5 with no assertions at all. Absence is a decision by default — make it a decision on purpose.
- **Orphan assertions.** A statement no one will own. Either find the owner or cut it; an unowned assertion will not be enforced regardless of what the document says.
- **Unreachable targets.** An enforceability target the organisation has no mechanism to reach. Lower the target or fund the mechanism, but do not publish a target that is decorative.

Phase 4 · Exceptions and escalation

A policy without a legitimate exception route is a policy people route around. Design the exception mechanism as part of the policy, not as an afterthought:

- **Who may grant an exception,** and must it differ by consequence class.
- **Every exception expires.** An exception with no expiry is an undocumented policy change. A fixed maximum duration, renewable by deliberate act, keeps the policy honest.
- **Exceptions are recorded against the assertion** they suspend, so coverage reporting can distinguish “not enforced” from “consciously excepted”.
- **Expiry re-raises the original condition,** rather than silently lapsing into permanent permission.

Phase 5 · Hand off to declaration

The output of this methodology is the input to the Declared Behavior Framework. The handoff should be mechanical:

01 Every system-classified assertion is issued with a stable reference so declared constraints can cite it.

02 Human-classified assertions are marked out of mechanised scope explicitly, so they are not later counted as coverage failures.

03 The assertion set is delivered in the coverage map's row format, so the first coverage assessment is a lookup rather than a decomposition exercise.

04 Baseline coverage is computed immediately — typically near zero, which is the correct and useful starting number.

> WHY THE BASELINE SHOULD BE MEASURED ON DAY ONE
> A client who sees “of your 34 system assertions, 0 are currently verifiable” on the day the policy is published understands exactly what has and has not been achieved. Publishing a policy without that number invites the assumption that writing it was the work. Establishing the baseline at publication converts the next engagement from a proposal into an obvious continuation.

5 · What a complete AI policy covers

Ten domains. Their value in an engagement is diagnostic: run a client's existing policy against this list and the empty rows are the finding. In practice most policies populate the first and the ninth and little else.

DOMAIN	THE QUESTION IT ANSWERS	TYPICAL STATE
Scope and definitions	What counts as an AI agent here, and what is out of scope	Usually present, usually too narrow to include autonomous action
Authorisation and registration	Who approves an agent, and must it be declared before running	Usually absent
Identity and accountability	Who owns each agent, and what happens when they leave	Almost always absent
Capability boundaries	What actions an agent may take	Sometimes present as prohibitions, rarely as permissions
Data handling	What data agents may access, and under what conditions	Often present but stated in unfalsifiable terms
Human oversight	What requires a person, and who may serve	Often present, rarely with independence conditions
Behavioral consistency	What constitutes acceptable change in an agent's behavior over time	Essentially always absent
Monitoring and evidence	What is recorded, retained, and for how long	Usually absent or delegated to “logging”
Acceptable use by staff	What people may do with AI tools	Almost always present; often the entire policy
Lifecycle and decommissioning	How agents are retired and credentials revoked	Almost always absent

The pattern worth showing a client. The domains organisations populate are the ones that govern people. The domains they leave empty are the ones that govern systems. That asymmetry is not carelessness — it is the natural result of writing an AI policy using the mental model of an acceptable-use policy, which is what almost everyone did first.

6 · Remediating an existing policy

Where a policy exists but cannot be enforced, do not start again. The client has invested in it, often visibly, and replacing it wholesale converts a sponsor into an opponent. Remediate in place.

01 **Decompose what exists.** Break the current policy into discrete assertions, one claim per row, without editing them. This is the coverage map's first step and it is done on the client's own words, which makes the output difficult to dispute.

02 **Classify every assertion.** System, human, or neither. Report the counts before reporting any opinion. The distribution is usually the most persuasive artifact in the engagement.

03 **Rewrite only the system assertions that fail the test.** Leave human assertions untouched — they are working as intended. This keeps the change surface small and the sponsor's investment intact.

04 **Fill the empty domains.** Use section 5 to identify what was never covered, then run phase 1 decisions for those domains only.

05 **Add owners and consequences to every retained assertion.** This is usually the highest-value low-effort change available, and it can be done without altering a single existing sentence.

06 **Publish the delta, not a new document.** An amended policy with a changelog is adopted; a replacement policy is re-litigated.

7 · Engagement shape and deliverables

Deliverables

- **Estate register.** What is running, what it touches, and what it would cost to be wrong.
- **Decision record.** The twelve decisions, as decided, with the dissent noted. This is the artifact that survives staff turnover.
- **Assertion set.** The policy in its atomic form, classified, owned, with consequences and enforceability targets.
- **Readable policy document.** Generated from the assertion set for board, staff and audit audiences.
- **Baseline coverage map.** The day-one measurement, in the Declared Behavior Framework's format.
- **Remediation plan.** Which assertions to mechanise first, sequenced by consequence rather than by ease.

Who must be in the room

A policy authored by one function fails in that function's characteristic way. The minimum viable group is four: someone who owns risk, someone who knows what the agents actually do, someone who can commit engineering effort, and someone with authority to decide when those three disagree. If the fourth is absent, the engagement produces recommendations rather than policy.

Positioning

This work is upstream of the assessment, not a substitute for it. The framing that holds:

“We will help you decide what should be true, in a form that can be checked. Then we will tell you whether it is.”

It is worth being explicit that the ideal client remains the one with a mature policy and no verification. This methodology exists so that a client who is not yet that becomes one — rather than being declined, or referred away and not returning.