

WHITE PAPER · SEPTEMBER 2026

What Assurance Actually Tests

A test plan for the AI Control Layer: populations, sampling, evidence, findings, and the limits of internal audit

Jermaine Leonard, AIGP

Principal, ArcPoint Consulting, LLC · Washington, DC
jleonard@arcpointconsulting.com

Certified Business Enterprise, District of Columbia, six designations · UEI JE2DLNX4BVQ3

This paper is analysis and general professional commentary. It is not an audit, attestation or compliance assessment of any organization, it does not establish auditing standards, and it is not legal advice.

Executive summary

An AI governance review that never selects a decision has audited the documentation of AI governance. It may have done that well. What it hasn't tested is whether anything in the organization would stop a consequential AI system that was producing bad outcomes, and that is the first thing an audit committee should want to know.

This is the third paper in ArcPoint's series on the AI Control Layer, the oversight structure that sits above AI systems and is defined by what it can stop, through four functions: Visibility, Validation, Control and Auditability.¹ The earlier papers argued that the layer needs assurance its operators can't overrule, and that human review of AI output works only when the reviewer has the information, the competence and time, the independence, and the exercised authority to disagree.² Neither said how an auditor would test any of that. This paper does. It is written for chief audit executives and audit committees, or their public-sector equivalents, and it matters to vendors too, because their logs, release notices and audit rights are becoming part of their clients' control environment.

The profession has little to work from. The IIA's AI Auditing Framework offers a checklist but no populations, sample sizes or evidence criteria.³ The method exists anyway. Two decades of Sarbanes-Oxley control testing worked out how to test design apart from operation, size a sample, rank evidence and grade a failure. It transfers after three adaptations: AI controls change without a change ticket, human review decays as volume rises, and the most important control in the layer has, in our experience, usually never been used.

For operating tests, the sampling unit is the consequential decision, drawn from a decision inventory, and not the AI system. Evidence the organization offers about how a control worked in the past counts only if the underlying record existed before the auditor asked for it and comes from a source the auditee cannot rewrite; the auditor's own re-performance and observation are a separate and often stronger category. A stop authority that has never been used gets its own conclusion, reported to the audit committee as a count. Internal audit can own most of the work, but it can't judge whether a validation was any good without a specialist who had no hand in building the system.

Appendix A is the test matrix: twenty-one tests across the four functions, each with its population, the evidence that satisfies it and the evidence that doesn't. It is also available from ArcPoint as an Excel workpaper, with result columns, a sample-size calculator and a summary for the audit committee pack. Government readers should take "audit committee" throughout to mean whoever receives internal audit's reports: an agency head, an inspector general, or an oversight board.

Method at a glance

The review moves in one direction:



An audit leader doesn't need a complete AI assurance program in the first review cycle. Section VIII sets out a first-year scope in five parts: an independent AI population, a decision inventory, deep tests of two or three high-consequence decisions (evidence, human review and stop authority), and one specialist validation review with independence from the builders written into the engagement.

That sequence gives the audit committee something it can govern: which decisions were tested, from what population, against what evidence standard, with what result.

I A diagnostic is not a test plan

Take one measurement every AI oversight program should own: the override rate, meaning the share of cases where human reviewers reject what the system recommends. If reviewers agree with the system on nearly every case, either the system is very good or the review is ceremonial, and the burden of proof sits with the first explanation. As a diagnostic, that works, as long as you remember that a better model and an easier case mix also push the rate down. Hand it to an auditor, though, and the questions start. Over what population, from what source, and is the log complete? What rate is too low, and how would you tell a sharper model from a duller review? If something is wrong, what's the finding, and how bad?

A diagnostic answers "is something wrong here?" A test answers "can I sign my name to a conclusion about this, and defend it to someone who has a reason to disagree?" A lot of AI assurance stops at the first question. Management fills in a questionnaire, the auditor reads it against a framework, inspects the policy, maybe walks through one system, and reports that the policy is incomplete. Section X looks at two published public-sector reports built largely that way. Every sentence in a report like that can be true while the audit committee still has no idea whether a consequential AI system could be stopped.

Audit practice has long settled where assurance sits: apart from the function it examines, where the people who run that function can't overrule it. What's missing is the Tuesday-morning detail of which records it pulls, how many, and what it concludes.

II What the profession has

Three bodies of guidance are available, and none of them is a test plan. The profession knows its coverage is thin: in the Internal Audit Foundation's Risk in Focus survey of more than 3,000 audit leaders, released in September 2026, 58 percent ranked digital disruption and AI among their top five organizational risks and 11 percent reported full audit coverage of it.⁴

The IIA's **Artificial Intelligence Auditing Framework** (2024) is candid that "it may be necessary to engage outside technical resources to assist with more technical aspects such as deciphering algorithms," and that "it is unlikely that internal auditors will ever have a mastery of knowledge on the topic."³ Its checklist says what to verify, not what to sample or what evidence passes.

GAO's AI Accountability Framework (GAO-21-519SP, June 2021) pairs each key practice with "procedures for auditors and third-party assessors."⁵ Most of them collect and review documents, and none specifies a population, a sample size or a pass standard. GAO is clear that the work needs "a multidisciplinary team that can address the technical aspects and the societal impacts of the AI system."⁵

ISO/IEC 42006:2025 sets requirements for bodies that certify AI management systems against ISO/IEC 42001.⁶ On BSI's description, it expects "technical expertise in data science, AI ethics and system validation in the auditing process."⁷ The IIA Standards set no AI-specific bar for internal auditors beyond the general rule that they "must possess or obtain the competencies to perform their responsibilities successfully."⁸

There is no IIA Topical Requirement for AI.⁹ The Third-Party requirement, in effect since September 15, 2026, applies where the audit function's risk assessment puts a vendor in scope. It never mentions AI, but its inventory and ongoing-monitoring controls should reach the AI a vendor supplies and a record of its model changes.⁹

What SOX already worked out

Sarbanes-Oxley produced a testing discipline, codified for external auditors in PCAOB AS 2201 and widely adopted by internal audit. It doesn't turn AI assurance into a financial-statement audit, but it answers most of the questions a diagnostic leaves open.

It tests design and operation separately. Operation asks "whether the control is operating as designed and whether the person performing the control possesses the necessary authority and competence to perform the control effectively."¹⁰ Authority and competence are two of the four oversight conditions, already written into the PCAOB's definition of a working control. A reviewer without the authority to reject the output, or the competence to disagree with it, would fail a SOX operating test as written.

It ranks evidence. AS 2201 lists tests "in order of the evidence that they ordinarily would produce, from least to most: inquiry, observation, inspection of relevant documentation, and re-performance of a control," and adds that "[i]nquiry alone does not provide sufficient evidence to support a conclusion about the effectiveness of a control."¹⁰ A governance review built on a management questionnaire is inquiry, however good the questions are.

It also scales evidence to risk, starts with a walkthrough that follows one transaction end to end (for AI, a consequential decision), and grades failures up to a material weakness, where there is a "reasonable possibility" that a material misstatement "will not be prevented or detected on a timely basis."¹⁰

Where the analogy needs adapting

Three things about AI controls require adaptation.

The object changes without a change ticket. AS 2201 lets auditors carry forward evidence about automated application controls into later years because "[e]ntirely automated application controls are generally not subject to breakdowns due to human failure," provided IT general controls are working and the auditor verifies the control "has not changed since the auditor established a baseline."¹¹ An AI system can behave differently with nothing new in the organization's change log. The vendor updates the model, the retrieval corpus grows, the inputs drift. Carrying forward last year's test is unsafe unless the change population includes events the organization didn't start.

Human review degrades with volume, by design. Manual controls suffer fatigue too, which is why volume is on the PCAOB's list of risk factors. Human review of AI output differs in two ways. The system is usually deployed to raise volume, so pressure on the reviewer is the goal, not a side effect. And the reviewer's default is the machine's answer, so fatigue shows up as rubber-stamping instead of visible error. Clinical medicine has measured the same habituation from the other direction. A review of 17 studies found drug-safety alerts in computerized physician order entry overridden "in 49% to 96% of cases," and traced much of it to alerts with low specificity.¹² Some of those overrides were justified, which is the point: a control that is usually wrong teaches people to clear it without reading. A review control tested in March can fail by September without any change to its design.

The most important control has usually never run. The stop test asks for the mechanism that can halt a system over its owner's objection, and when it was last used.¹ In our experience, the answer to the second half is usually never. SOX practice already handles controls with no occurrences: the tester notes that none occurred and relies on design. That's fine for a low-frequency reconciliation. It isn't fine when the untested control is the one the whole layer depends on, and Section VII proposes reporting it differently.

IV The sampling unit is the decision

A system inventory counts exposure. A decision inventory counts authority. Twenty tools that draft text fill twenty rows on the first and none on the second, while one system that sets a price, screens an applicant or releases a payment is the only row that matters on the second.² For assurance, that difference becomes the sampling frame. A review that samples systems spends its hours on the drafting tools, because there are more of them. A review that samples decisions spends its hours where authority moved.

The review works from five populations.

The inventory. The first test is completeness, and since management's inventory is what's under test, it can't be the source of the sample. The auditor builds an independent population from records the governance team doesn't keep: accounts-payable and procurement records for AI vendors and AI-bearing products, single sign-on and OAuth grants, proxy logs showing traffic to model endpoints, and release notes in which existing vendors announce AI features. A sample from that population is traced back to the inventory, and anything missing is an exception. This is the test that finds embedded AI, which in our experience is where most unreported AI lives.

Consequential decisions. From the decision inventory, the auditor picks decisions that are both consequential and hard to reverse, then samples individual instances from the system's own records. Each is traced end to end, from what the system produced to what happened next, and checked against the record.

Human reviews. Where a decision has a human-review step, the population is the set of reviewed cases. Attributes that need judgment, such as whether the reviewer saw the reasons for the output, are tested on a sample. Arithmetic attributes (override rate, cases per reviewer per day, time to decision) should be computed across every case. The log is the population, and computing a rate over all of it costs about the same as computing it over sixty. When the override rate moves, two further tests tell a better model from a weaker review. With management's written agreement, seed a small number of fictitious cases carrying known-wrong recommendations into the live queue and count how many reviewers catch. Seeded testing needs legal approval, operational safeguards and no effect on any real applicant, beneficiary, employee or other affected person. Where seeding isn't acceptable, have a qualified reviewer independently re-decide a random sample and compare. Auditors have used test transactions this way for decades.

Changes. New model versions, prompt and configuration edits, threshold changes, retrieval-corpus updates and vendor model updates. Each sampled change is traced to validation evidence dated before it reached production. Vendor-initiated changes have to be in the population, which means someone has to be collecting them. If no one is, that is the finding.

Interventions. Every suspension, rollback or shutdown at the system level, as opposed to an override on one case. The population is small, so test all of it. If it's empty, see Section VII.

How many to pull

Sample sizes for control tests come from attribute sampling: how many items must you examine, finding no more than the allowed number of failures, to say with stated confidence that the true failure rate is below a tolerable limit?

Confidence	Tolerable failure rate	Sample needed, zero failures allowed	Sample needed, one failure allowed
95%	5%	59	93
90%	5%	45	77
95%	10%	29	46
90%	10%	22	38

Computed from the binomial distribution: the smallest sample for which the probability of observing no more than the stated number of failures, if the true rate equals the tolerable rate, falls to or below one minus the confidence level.

Audit firms' SOX tables rest on the same arithmetic, calibrated to how often a control runs, and AI decision populations nearly always sit at the high-frequency end. Decide in advance what one exception means. One failure in 59 doesn't prove the control failed, but the auditor then has to expand to 93 or conclude it isn't effective at that confidence. For irreversible decisions, the tolerable rate should be low and the confidence high. A 10 percent tolerable failure rate on a control over hiring decisions is something the audit committee should be asked to endorse in so many words.

Judgmental samples have a place and a cost. The New York State Comptroller's AI governance audit, discussed in Section X, said plainly that "because we used a non-statistical sampling approach for our testing, we cannot project the results to the respective populations."¹³ Audit committees should always know which kind of sample they're reading about.

V What counts as evidence

The standard for auditability fits in one line: if the evidence is generated by the request for evidence, what exists is a reconstruction.¹ For an auditor, that line becomes a test, and it may be the most useful single test in the matrix.

Call it the timestamp test. It applies to evidence the organization offers about how a control worked in the past. For each item, establish when the underlying record was created relative to the request, using a date the auditee can't edit: a ticketing system's audit trail, a document-management version history, a log with restricted write access. File metadata isn't enough. System logs, change tickets, validation reports dated before release, review records made at the moment of review, and minutes that record a decision rather than a discussion all existed before the auditor asked. That is operating evidence. A report exported during fieldwork can still be strong evidence if it is a query against pre-existing records the auditee can't alter and the auditor tests the source. A memo explaining how the control works, a configuration screenshot taken during fieldwork, a spreadsheet compiled from memory, or a live demonstration was made after the request. At best it shows design, and at worst it's a reconstruction.

Evidence the auditor generates is a different matter. Re-performing a control, watching it run, or querying production records all happen after the request by definition, and AS 2201 ranks re-performance as the strongest test available.¹⁰ The timestamp test leaves those alone. What it rules out is the organization assembling its own past on demand.

For evidence the organization supplies, that gives a rough order of strength, consistent with AS 2201: system records created in the ordinary course of operation, with their own integrity controls; then contemporaneous human records, such as a validation report signed before go-live; then documents describing the control, which establish design and nothing about operation; then demonstrations, performed for the auditor on a day the auditee picks; and inquiry last.

AI audits lean on demonstrations, which are persuasive to watch and weak as evidence: a chosen input, at a chosen time, run by someone with a commercial stake in the result. In the New York audit, auditors "viewed vendor demonstrations and interviewed the vendor about the underlying technology" for each sampled use case. In one case the vendor first said the system used AI and later "clarified

that DOT's use case does not use AI." The auditors concluded "there is insufficient evidence to determine whether this use case used AI, as the vendor has the capability to use AI and does so for other clients."¹³ That was the right call, and it shows the limit of demonstration and inquiry: neither can establish whether a system in production does what its operator says.

The log itself needs testing too. SOX practice tests information produced by the entity before relying on it, and AI records are no different. Before computing an override rate from a review log, reconcile its record count to an independent count of decisions for the period, and confirm that reviewers can't edit their own entries. A log that leaves out the cases reviewers skipped will produce a perfect override rate.

VI Testing the four functions

The layer's four functions are Visibility, Validation, Control and Auditability. They aren't maturity stages. An organization has each one or it doesn't, and the weakest one sets what the layer can do.¹ So a review that tests Visibility thoroughly and skips Control has produced an opinion on Visibility, not a partial opinion on the layer.

Appendix A sets out all twenty-one tests. The ones that decide each function are these.

Visibility turns on inventory completeness (V-1) and on whether a decision inventory records, for each consequential decision, how much authority was delegated to the system and whether the action can be reversed without a human (V-2). V-2 is a design test followed by a sample: take consequential decisions from process documents the governance team didn't write, and check each against the decision inventory. The usual failure is a system inventory with no decision rows at all, which should be reported as a design deficiency.

Validation turns on whether each release of a consequential system was tested before deployment, against criteria written in advance, by someone other than the builders (Va-1, Va-2), and whether changes, including vendor changes, triggered revalidation (Va-3). The timestamp test carries most of the load: criteria should predate results, and results should predate release. Criteria written after the results just describe what the system happened to do. Whether the criteria were adequate (Va-5) is the part internal audit can't judge alone.

Control has the most tests. The stop authority (C-1, C-2) is tested for design by confirming that a named person outside the owning team holds it, with written triggers and a mechanism that exists in the system and not just in a memo, and for operation against the intervention population. The four oversight conditions become attributes tested on each sampled review:

- *Information* (C-3): rebuild what the reviewer saw at the moment of review. If the reasons behind the output weren't on screen, the condition fails for that case, whatever the model card says.
- *Competence and time* (C-4): check the reviewer's role and training against the requirement, and compute cases per reviewer per day and time per case across the population. The process owner should set acceptable limits before the test, and the auditor should challenge them.

- *Independence* (C-5): trace the reviewer's reporting line and performance measures. A reviewer measured on throughput fails whatever the reporting line.
- *Authority, exercised* (C-6): compute the override rate across all cases, by reviewer and by month, and find out who owns and reports the number. A rate that falls as volume rises is a flag that has to be resolved, and it is resolved by seeded cases or independent re-review, not by management's explanation.

Auditability turns on the timestamp test (A-1), log integrity (A-2) and reconstruction (A-3). For A-3, take a sampled decision from several months back and rebuild it end to end (inputs, model version, output, reviewer, action, outcome) using only records that existed at the time. Two hours is a fair benchmark; report the actual time.

DC agencies and their vendors have one more test (A-4). The District's standard AI contract clause E. 4 gives the District the right to observe or audit relevant contractor work processes at no additional cost.¹ A contract right that has never been used looks, at audit, like one that doesn't exist, and whether this one has been used is a documented yes-or-no question.

VII What a finding looks like

Government Auditing Standards require that "when auditors identify findings, they should plan and perform procedures to develop the criteria, condition, cause, and effect of the findings to the extent that these elements are relevant and necessary to achieve the audit objectives." Criteria "identify the required or desired state or expectation"; condition "is a situation that exists"; cause is "the factor or factors responsible for the difference between the condition and the criteria"; effect is "the outcome or consequence resulting from the difference."¹⁴ It forces precision where AI findings tend to go vague: the criteria have to be a specific requirement, not a framework, and the effect has to name a consequence, not a risk category.

Here is an illustrative finding. It is a composite and describes no actual engagement.

Finding C-3/C-4/C-6. Human review of benefit-eligibility recommendations is not operating as designed.

Criteria. The agency's AI use procedure requires that each recommendation to deny eligibility be reviewed by a qualified caseworker who may accept or reject it (Procedure §4.2(a)), and that the review screen display the factors behind each recommendation (§4.2(b)). The agency's stated oversight standard requires that reviewers have sufficient time per case to evaluate the reasons given (§4.3).

Condition. We tested 59 denial recommendations selected at random from 41,380 made between January and June. In 14, the reviewer's screen did not display the factors behind the recommendation; the record showed only the score. Across the full population, the reviewer override rate fell from 6.1 percent in January to 0.4 percent in June while monthly volume rose 2.3 times. Median time from presentation to decision fell from 3 minutes 40 seconds to 38 seconds. An independent caseworker re-decided 40 of the sampled denials and would have approved 9. No one in the program office receives or reviews any of these figures.

Cause. A May interface change removed the factors panel to reduce load time. It was treated as a user-interface change and not routed for validation. Staffing was not adjusted for volume growth.

Effect. For at least part of the period, denials were determined by the system with a caseworker's approval recorded. Denials are reversible only on appeal, which requires the applicant to act.

Rating. Grade 1, reportable to the audit committee. Fourteen exceptions in 59 far exceed any tolerable rate, independent re-review shows the failure is producing wrong denials, no compensating control was identified, and a denial can be reversed only if the applicant appeals.

Most AI governance findings don't do what this one does. It names a requirement for each exception, pairs a sample with whole-population arithmetic, confirms harm with independent re-review, and traces the cause to a change that went around validation, which gives management something to fix besides "strengthen governance."

Grading severity: reversibility and likelihood

SOX grades a deficiency on how large the misstatement could be and whether there's a "reasonable possibility" it goes undetected.¹⁰ The model below is ArcPoint's proposed severity scale for AI control findings, adapted from that logic. Magnitude comes from consequence and reversibility: whether a person's access to money, work, benefits, housing, safety or legal standing turns on the decision, and whether the outcome can be undone without that person having to act. Likelihood comes from the evidence: the exception rate against the tolerable rate, the trend in the whole-population metrics, and whether a compensating control would catch the failure first.

Grade 1, the analogue of a material weakness, covers a consequential decision that can't be reversed without the affected person's action, where either a control failure carries a reasonable possibility of harm going undetected or a control that prevents or stops harm is missing altogether: no stop authority, or no human review where the organization's own rules require one. A decision missing from the inventory is Grade 2, unless the completeness test finds a consequential decision nobody in the organization knew about, with no owner and no review. That is Grade 1, because nobody is positioned to stop it. Grade 1 goes to the audit committee with a named owner and a remediation date. **Grade 2**,

the analogue of a significant deficiency, covers a failure over a consequential but reversible decision, a design gap that doesn't remove a required control, or an isolated exception over an irreversible decision that a compensating control would have caught. **Grade 3** is everything else, reported to management.

Two conclusions sit outside that scale. **Unevidenced** means the control may work, but the only evidence offered failed the timestamp test, so the auditor can't call it effective. **Never operated** applies to a control with an empty population. SOX testers already record "no instances occurred"; what should change is where it's reported. For a stop authority, the auditor can test design, walk through the mechanism and ideally watch a recorded exercise outside production, but can't conclude that it operates effectively, and should say so in those words. The number of consequential systems whose stop authority has never been used belongs on the audit committee's page as a headline figure, not in a workpaper footnote.

VIII Scoping the first year

A first-year review that tries to cover every AI system will cover none of them properly. The decision inventory is the scoping document. If it doesn't exist, building the population for it is the first year's work.

A defensible first-year scope looks like this:

- **Organization-wide inventory completeness** (V-1, V-4), the only organization-wide test in year one, because every later test depends on it.
- **Decision inventory existence and design** (V-2, V-3). If there isn't one, that's a design finding, and the audit team drafts its own scoping population.
- **Deep testing of two or three consequential decisions**, chosen for consequence, irreversibility and volume together, each with the full Control and Auditability test set. In DC, Mayor's Order 2024-028 §III attaches a documented obligation to every agency deployment, so deployment records are a ready-made source.¹
- **The stop test for each selected decision** (C-1, C-2): design, walkthrough, and a never-operated conclusion if that's what the evidence supports.
- **One validation deep dive with a specialist** (Va-5), on the highest-consequence system, enough to learn what the organization's validations contain.

Years two and three rotate the deep tests across the rest of the decision inventory, rerun completeness, and follow up every Grade 1 finding. Don't carry forward prior-year evidence for any AI control unless the change population, vendor changes included, has been tested and found clean.

Timing matters. A review that isn't in the annual audit plan won't happen, and for calendar-year organizations the plan is usually approved late in the year before. DC agencies have an extra date: the AI Taskforce and Advisory Group sunset on December 31, 2026, while the per-deployment mandate does not.¹ From January 1, the capacity to produce and check the evidence for each deployment is either inside the agency or bought under contract.

IX What internal audit can and cannot assess alone

The IIA framework is right that internal auditors don't need to be AI experts to audit AI, and that some of the work needs outside help.³ Internal audit can own fifteen of the twenty-one tests outright with ordinary IT audit skills, and gather the evidence for most of the other six. It needs specialist help to judge whether reviewers are shown enough to disagree with the output, whether acceptance criteria and drift thresholds are set at meaningful levels, whether revalidation covered what a change could affect, and whether a log's integrity controls would catch tampering.

What it shouldn't attempt alone is judging whether a validation was adequate: whether the acceptance criteria measured the right properties, whether bias and fairness testing used appropriate methods and populations, whether robustness and adversarial testing (including prompt injection for agentic systems) meant anything, and re-performing any of it. The IIA Standards let internal auditors "obtain" competencies rather than possess them.⁸ In practice, obtaining means bringing in a specialist.

Specialist work counts as assurance only on two conditions. The first is independence from the builders. AS 2201's logic on using others' work applies: the auditor "should assess the competence and objectivity of the persons whose work the auditor plans to use," and "as the risk associated with a control increases, the need for the auditor to perform his or her own work on the control increases."¹⁰ Neither the firm that built or implemented the system nor the internal team that built it is independent of it, whatever the engagement letter says. Independence should be a written exclusion checked at engagement, not a disclosure. Federal agencies already work to this idea: OMB M-25-21 requires the impact assessment for high-impact AI to include "results of independent review," from a reviewer inside the agency who wasn't involved in developing the system.¹⁵ The second condition is that the specialist works to the audit's test plan and criteria. A report concluding that a model is "robust" is a design opinion unless it says what was tested, against which criteria set in advance, on which population, with what result.

X Case study: the ceiling of document-and-interview assurance

New York's Office of the State Comptroller has published two AI governance reports that show what document review, interviews and demonstrations can reach. The statewide audit (Report 2023-S-50, April 2025), conducted under generally accepted government auditing standards, sampled four agencies and one use case each, relying on interviews, supporting documents, and vendor demonstrations and interviews "about the underlying technology."¹³ Its best moment was a classification test: the Department of Motor Vehicles had left its facial recognition system out of its AI policies, and the auditors checked that call against the state's own definition and found the system met it.¹³ That's the inventory completeness test, done once, by someone independent of the person who made the call.

Most of its principal findings, though, were findings of absence, such as no testing procedures and no formal inventory. Absence is the cheapest thing an audit can establish, because it needs only a request and a null response. An agency that had built an inventory and test procedures would have passed, and the audit couldn't then have said whether any of it worked. Nothing in the stated method selects a decision, examines a review or tests a log.

The New York City follow-up (Report 2025-F-17, August 27, 2026) found that the city's Office of Technology and Innovation "still relies on agencies to self-report and does not maintain an inventory of AI tools and systems used by the City."¹⁶ Its stated objective was "to assess the extent of implementation, as of April 13, 2026, of the three recommendations included in our initial audit report," so it would be unfair to expect a completeness test.¹⁶ But the gap is plain. Assurance established that nobody was looking for unreported AI. The next step is for assurance to look.

The Comptroller's auditors worked within their objectives and wrote clearly; the limit is the method's. Document-and-interview assurance suits organizations that haven't built anything yet. Once they build, assurance has to test whether controls work, not just whether they exist.

XI The case for waiting

The case against putting this in next year's plan: there's no AI Topical Requirement, ISO/IEC 42006 is barely a year old, the NIST AI RMF is being revised, the EU just pushed its high-risk obligations to December 2027,² most audit functions lack the skills, and vendors won't expose everything. Better to audit at the policy level for another year or two and wait for the profession to publish a method. We sell AI governance work, so weigh our answer accordingly.

The method was published twenty years ago for a different object, and a Topical Requirement would add the obligation to use it, not the ability. Vendor opacity is a reason to test audit rights, release notices, logging and validation artifacts as controls, not a reason to skip the plan. Waiting also has a cost. Threshold drift, collapsing override rates and vendor changes that skip validation are invisible to a policy-level review and get worse over time, building up behind a report that says the program was reviewed. That is the audit function's own version of what Madeleine Clare Elish called a moral crumple zone, where responsibility lands on a participant with limited control over the outcome.¹⁷

One concession: if nothing an organization runs with AI influences a consequential decision, a first-year review comes down to an inventory completeness test and a walkthrough of whoever can switch things off. It's still worth doing, because the completeness test is how an organization finds out whether that belief is true.

XII For the audit committee

An audit committee doesn't need to understand model validation to hold an AI assurance plan to account. It needs questions whose answers can be checked.

What was the sampling unit, systems or decisions? How was the inventory tested for completeness, and from which sources outside the governance team? How many consequential decisions were tested, from what population, at what confidence? How much of the evidence the organization supplied existed before the request? How many stop authorities have never been used? Who reviewed the validations, and what's their relationship to the people who built the system?

A report that can answer those has tested something. A report that can't has read something.

At least one buyer has started checking for itself: the District of Columbia has written audit rights into its standard AI contract clauses. Certification auditors now have a standard of their own. The open question is whether the assurance function inside the organization gets there first, with a test plan of its own.

A Test matrix

Functions follow the AI Control Layer definition. **Pop.** = population to sample from. **Spec.** = specialist needed (— none; ● assists; ●● required for the conclusion). Tests marked † use full-population analysis rather than, or in addition to, sampling. Sample sizes follow Section IV; where the population is under about 25 items, test all.

Visibility

Ref	Control objective	Test of design	Test of operation	Pop.	Evidence that satisfies	Evidence that does not	Spec.
V-1	The AI inventory is complete	Inventory process defines AI, covers embedded and unsanctioned AI, has an owner and a refresh cadence	Build an independent population from AP/ procurement, SSO/OAuth grants, network logs to model endpoints, vendor release notes; trace a sample to the inventory (confirm the legal basis for log review first)	Independent population, not the inventory	Traced items found in the inventory with matching owner; exceptions logged	Management attestation of completeness; an inventory built from a survey with no outward-in check	—
V-2	Each consequential decision is inventoried with delegated authority level and reversibility	Decision inventory exists; fields include authority level (inform / recommend / decide / act), reversibility, affected party, intervening role	Select consequential decisions from process documentation not written by the governance team; trace to the decision inventory	Process documentation	Each traced decision present with accurate authority level and reversibility	A system inventory with no decision rows; rows that list systems as "decisions"	—
V-3	Each consequential decision has a named business owner	Ownership field populated with a person, not a committee or team	For sampled decisions, confirm the owner knows they own it and can describe the stop trigger	V-2 rows	Named owner confirms and describes; owner appears in change and incident records	A team name; an owner who learns of the role at interview	—
V-4	AI embedded in third-party products is identified	Vendor onboarding and renewal ask about AI features and model changes	Sample in-scope vendors; compare vendor release notes and contracts to inventory	Vendor master, in-scope vendors	Embedded AI recorded before the audit request	Vendor questionnaire answered after the request	—

Validation

Ref	Control objective	Test of design	Test of operation	Pop.	Evidence that satisfies	Evidence that does not	Spec.
Va-1	Consequential systems are validated before deployment against criteria set in advance	Procedure requires written acceptance criteria before testing	For each release in scope, compare dates: criteria < results < release	Releases of in-scope systems	Dated criteria, dated results, approval before release	Criteria dated after results; a vendor benchmark in place of a deployment validation	●
Va-2	Validation is performed or reviewed by someone independent of the builders	Procedure names who validates and their reporting line	For sampled validations, trace validator's reporting line and role in build	Va-1 sample	Validator outside the build team's reporting line	Builder's self-test signed by builder's manager	—
Va-3	Changes trigger revalidation, including vendor-initiated changes	Change policy defines which changes require revalidation; vendor model updates captured	Trace sampled changes (model, prompt, threshold, corpus, vendor update) to revalidation before production	Change population incl. vendor notices	Revalidation dated before change reached production	No record of vendor changes; UI changes excluded by default	●
Va-4	Ongoing monitoring detects drift and performance breaches and triggers action	Metrics, thresholds and responders defined	† Analyse monitoring output for the period; sample breaches to action taken	Monitoring alerts	Breach → ticket → action, all timestamped	Dashboards with no thresholds; alerts with no owner	●

Ref	Control objective	Test of design	Test of operation	Pop.	Evidence that satisfies	Evidence that does not	Spec.
Va-5	Validation is adequate: right properties, methods and populations, incl. bias/fairness and robustness	Criteria cover accuracy, fairness across relevant groups, robustness, security (incl. prompt injection for agentic systems)	Specialist evaluates and, where warranted, re-performs a selected validation	Highest-consequence system (year one: one)	Specialist report stating criteria, method, population, results	A validation summary with no method; an opinion that the model is "robust"	●●

Control

Ref	Control objective	Test of design	Test of operation	Pop.	Evidence that satisfies	Evidence that does not	Spec.
C-1	A named person outside the owning team can halt each consequential system	Authority documented with triggers; mechanism exists in the system	Walkthrough of the mechanism with the authority holder	In-scope decisions	Named holder, written triggers, technical mechanism observed	Authority vested in the product owner; "we can turn it off" with no trigger	—
C-2	The stop authority operates	—	Test every intervention; if none, observe a recorded exercise in non-production and report never operated	Interventions (usually small or empty)	Intervention records with trigger, decision-maker, time to halt, outcome	Absence of interventions reported as effectiveness	—
C-3	Reviewers have the information to disagree (oversight condition 1)	Review interface specified to show reasons/factors, not only a score	For sampled reviews, reconstruct what was on screen at the moment of review	Reviewed cases	Interface record or version history showing factors present at that time	Current screenshot; model card	●
C-4	Reviewers have competence and time (condition 2)	Role, training and case-time standards defined in advance by process owner	Check reviewer role/training for sample; † compute cases/reviewer/day and time per case	Reviewed cases; reviewer roster	Metrics within defined limits across the period	No defined limits; limits set after metrics seen	—
C-5	Reviewers are independent of the output under review (condition 3)	Reporting lines and performance measures documented	Trace sampled reviewers' reporting line and goals	Reviewer roster	Reviewer not measured on the throughput the system raises	Reviewers in the owning team with throughput targets	—
C-6	Reviewer authority is exercised and monitored (condition 4)	Override rate defined, owned and reported	† Compute override rate by reviewer and month against volume; confirm who receives it; resolve any fall with safeguarded seeded cases (with written agreement) or independent re-review of a sample	Reviewed cases (full)	Stable rate, or a fall shown by seeded cases or re-review to reflect a better model; named recipient; action on anomalies	No owner for the metric; a falling rate explained by management without testing	—
C-7	Thresholds that route cases to humans are governed	Who may change thresholds is defined; changes require approval	Trace sampled threshold changes to approval and revalidation (Va-3)	Threshold change log	Approved, dated, revalidated	Thresholds changed by the operating team without approval	—
C-8	Policy binds at an enforcement point	Prohibited uses and data restrictions enforced technically where possible (blocking, access, DLP)	With written authorization, attempt a prohibited action, or inspect logs of attempts	Enforcement logs	Blocks recorded in normal operation	Policy text with no enforcing mechanism	—

Auditability

Ref	Control objective	Test of design	Test of operation	Pop.	Evidence that satisfies	Evidence that does not	Spec.
A-1	Evidence is produced in operation (timestamp test)	Record-keeping requirements defined per consequential decision	For every item of auditee-provided evidence relied on in this review, compare the underlying record's creation date to request date	All evidence obtained	Created before the request, per a system-generated date the auditee can't alter	Memos, screenshots, spreadsheets and demonstrations produced for the audit; file-metadata dates	—
A-2	Logs are complete and cannot be altered by those they record	Log retention and access controls defined	† Reconcile log record counts to independent decision counts; test edit rights	Decision logs	Counts reconcile; write access restricted	Log missing skipped cases; reviewers able to edit entries	●
A-3	A decision can be reconstructed end to end	Retention covers inputs, model version, output, reviewer, action, outcome	Reconstruct sampled historic decisions using only contemporaneous records; record time taken	Consequential decisions, several months old	Full reconstruction within a set limit (two hours as a benchmark)	Reconstruction that needs interviews to fill gaps	—
A-4	Contractual audit and reporting rights are exercised	AI contracts carry audit/reporting rights (for DC, clause E.4)	Confirm rights exercised in the period and results acted on	In-scope AI contracts	Dated exercise of the right and follow-up	Right present, never invoked	—

About ArcPoint

ArcPoint Consulting is a senior-led AI governance, risk, and control advisory practice in Washington, DC. We do not help organizations adopt AI. We help them govern it.

We build the AI Control Layer: the evidence, oversight, and documentation that make an AI deployment visible, validated, controlled, and auditable. It is produced by ArcPoint's Weftline Framework, which assesses five dimensions together: Strategic Intent, Decision Authority, Operational Integration, Workforce Enablement, and Technical Infrastructure, with AI Influence Mapping running across all five. The framework builds on the NIST AI Risk Management Framework and adds structured ethical value elicitation drawn from IEEE 7000 along with a defined bias-testing protocol. Jurisdiction-specific requirements layer on top.

When we verify an AI system independently, we never verify one we built or implemented, and for 24 months after implementation work for an organization we verify its systems only where they are demonstrably unrelated to that work.

Every engagement is delivered and signed by the principal.

Jermaine Leonard, AIGP · ArcPoint Consulting, LLC · Washington, DC ·
jleonard@arcpointconsulting.com

Certified Business Enterprise, District of Columbia, six designations · UEI JE2DLNX4BVQ3

This paper is analysis and general professional commentary. It is not an audit, attestation or compliance assessment of any organization, it does not establish auditing standards, and it is not legal advice. The illustrative finding in Section VII is a composite and describes no actual organization.

Endnotes

1. ArcPoint Consulting, *The AI Control Layer: Why every AI-enabled organization ends up building one, and what it costs to build it late*, August 2026. Source of the layer's definition (the stop test, the four functions, the separation requirement, and the statement that the functions are not maturity stages), the auditability standard restated in Section V, and the DC material: AI Procurement Handbook clause E.4 audit rights at no additional cost (paraphrased there and here, not quoted), Mayor's Order 2024-028 §III per-deployment duty, and the §IV.J/§VI.J December 31, 2026 sunset. Each was verified against primary sources in that paper's sources and verification table.

2. ArcPoint Consulting, *AI Governance as Organizational Design: Why the next era of AI governance is an operating model challenge*, August 2026. Source of the four oversight conditions, the override-rate measure, the system/decision inventory distinction, and the reversibility boundary. Also the verified source (2026-08-22) for the Regulation (EU) 2026/1744 deferral of Annex III obligations to December 2, 2027 and the NIST AI RMF revision under the July 2025 AI Action Plan, both mentioned in Section XI; both re-checked live 2026-09-28 against EUR-Lex (<https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng>) and nist.gov (<https://www.nist.gov/itl/ai-risk-management-framework>) and unchanged.

3. The Institute of Internal Auditors, *The IIA's Artificial Intelligence Auditing Framework* (2024 update). Quotations: "Internal auditors are not expected to be experts on every audit topic..."; "It may be necessary to engage outside technical resources to assist with more technical aspects such as deciphering algorithms"; "Because AI is so highly complex and changing, it is unlikely that internal auditors will ever have a mastery of knowledge on the topic; think of auditing AI as a progression, not as a destination." <https://www.theiia.org/globalassets/site/content/tools/professional/aiframework-sept-2024-update.pdf> — read 2026-09-27.

4. The Institute of Internal Auditors / Internal Audit Foundation, *Risk in Focus: Global* (2026), more than 3,000 audit leaders surveyed. Anthony Pugliese, "Risk in Focus: Top 10 Takeaways from This Year's Findings," *Internal Auditor* (IIA), September 16, 2026: digital disruption, including AI, rose "ten percentage points to 58% globally" as a top-five risk; audit priority rose "ten percentage points to 42%"; "Just 23% of respondents rate governance of digital disruption as managed or optimized"; "Only 11% of respondents report full audit coverage for digital disruption." <https://internalauditor.theiia.org/en/voices/2026/september/risk-in-focus-top-10-takeaways-from-this-years-findings/> — read 2026-09-27. The IIA press release as distributed phrases the last figure as "just 11% considered internal audit coverage fully adequate" (<https://www.einpresswire.com/article/942907666/internal-auditors-shift-focus-toward-ai-and-other-emerging-risks>, read 2026-09-27). Full report not read.
5. U.S. Government Accountability Office, *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities*, GAO-21-519SP, June 30, 2021. Summary page (<https://www.gao.gov/products/gao-21-519sp>) for the practice/question/procedure structure; report PDF (<https://www.gao.gov/assets/gao-21-519sp.pdf>) p. 12 for audit procedures "developed by reviewing the types of evidence noted in the Government Auditing Standards," p. 37 for the Performance procedure quoted, and p. 27 for the multidisciplinary-team statement — read 2026-09-27.
6. ISO/IEC 42006:2025, *Information technology — Artificial intelligence — Requirements for bodies providing audit and certification of artificial intelligence management systems*, edition 1, July 2025; supplements ISO/IEC 17021-1. <https://www.iso.org/standard/42006> — read 2026-09-27. The standard's clauses were not read; this paper asserts nothing about them beyond ISO's published description.
7. Gavin Jones (BSI), "New benchmark for bodies certifying AI management systems," Chartered Quality Institute, July 10, 2025: the standard's competency framework "provides AI auditors with defined expectations around skills, ensuring technical expertise in data science, AI ethics and system validation in the auditing process." <https://www.quality.org/article/new-benchmark-bodies-certifying-ai-management-systems> — read 2026-09-27. Attributed commentary, not the standard's text.
8. The Institute of Internal Auditors, *Global Internal Audit Standards*, published January 9, 2024, Standard 3.1 Competency: "Internal auditors must possess or obtain the competencies to perform their responsibilities successfully." https://www.theiia.org/globalassets/site/standards/globalinternalauditstandards_2024january9.pdf — read 2026-09-27.
9. The Institute of Internal Auditors, *Topical Requirements*. Applicability: "Depending on the results of the internal audit function's risk assessment, internal auditors must apply Topical Requirements in conformance with the Global Internal Audit Standards when providing assurance services on the topic." Effective dates: Cybersecurity February 5, 2026; Third-Party September 15, 2026; Organizational Behavior December 15, 2026; Organizational Resilience April 30, 2027. In development: Talent Management, Anti-Corruption. No AI Topical Requirement listed. <https://www.theiia.org/en/standards/2024-standards/topical-requirements/> — read 2026-09-27. Third-Party Topical Requirement text: defines a third party as "an external individual, group, or entity with whom an organization establishes a business relationship to obtain products or services," including outsourced service providers; its controls include maintaining a third-party inventory and ongoing monitoring; it contains no reference to AI. https://www.theiia.org/globalassets/site/standards/topical-requirements/third-party/third_party_topical_requirement_english.pdf — read 2026-09-27.
10. Public Company Accounting Oversight Board, AS 2201, *An Audit of Internal Control Over Financial Reporting That Is Integrated with An Audit of Financial Statements*: ¶1.18–.19 (using the work of others), ¶1.37 (walkthroughs), ¶1.42 (design effectiveness), ¶1.44 (operating effectiveness), ¶1.45 (procedures), ¶1.46–.47 (risk and evidence; risk factors), ¶1.50 and its Note (ordering of tests from least to most evidence; inquiry alone insufficient), Appendix A ¶1.A3, .A7, .A11 (deficiency, material weakness, significant deficiency). <https://pcaobus.org/oversight/standards/auditing-standards/details/AS2201> — read 2026-09-27. A version amended by PCAOB Release No. 2024-005 takes effect December 15, 2026; the wording of ¶1.44 and the ¶1.50 Note quoted here is unchanged in that version (checked 2026-09-27).
11. AS 2201 ¶1.60 ("The auditor may also use a benchmarking strategy for automated application controls in subsequent years' audits") and Appendix B ¶1.B28: "Entirely automated application controls are generally not subject to breakdowns due to human failure. This feature allows the auditor to use a 'benchmarking' strategy." ¶1.B29 conditions it on effective general controls over program changes, access and computer operations, and on the auditor verifying that the control "has not changed since the auditor established a baseline." Appendix B text read from the PCAOB's page for Auditing Standard No. 5, Appendix B (https://pcaobus.org/oversight/standards/archived-standards/pre-reorganized-auditing-standards-interpretations/details/Auditing_Standard_5_Appendix_B), carried into AS 2201 as ¶1.B28–.B29 — read 2026-09-27.
12. Heleen van der Sijs, Jos Aarts, Arnold Vulto and Marc Berg, "Overriding of Drug Safety Alerts in Computerized Physician Order Entry," *Journal of the American Medical Informatics Association* 13(2), 2006, 138–147. A literature review of 17 papers; abstract: "Drug safety alerts are overridden by clinicians in 49% to 96% of cases. Alert overriding may often be justified..." and names "low specificity" among the error-producing conditions. <https://academic.oup.com/jamia/article-abstract/13/2/138/729701>; abstract read via Europe PMC (PMID 16357358), 2026-09-27.

13. Office of the New York State Comptroller, *New York State Artificial Intelligence Governance*, Report 2023-S-50, April 2025. Quotations from pp. 1, 8, 15–16, 19–23; GAGAS statement p. 24. <https://www.osc.ny.gov/files/state-agencies/audits/pdf/sga-2025-23s50.pdf> — read 2026-09-27.

14. U.S. Government Accountability Office, *Government Auditing Standards: 2024 Revision*, GAO-24-106786, February 2024; applies to performance audits beginning on or after December 15, 2025. ¶8.116 (requirement to develop criteria, condition, cause and effect), ¶8.124–8.127 (definitions of each element), pp. 214–216. The glossary definition of "finding" cross-references the same elements at ¶¶6.17, 6.19, 7.19, 7.21, 8.116 and 8.118. <https://www.gao.gov/assets/d24106786.pdf> — read 2026-09-27 from the full PDF supplied by the owner.

15. Office of Management and Budget, Memorandum M-25-21, "Accelerating Federal Use of AI through Innovation, Governance, and Public Trust," April 3, 2025, §4(b)(ii): the AI impact assessment for a high-impact use case must include "results of independent review," supported by an independent reviewer within the agency not involved in the system's development. <https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf> — read 2026-09-28.

16. Office of the New York State Comptroller, *Artificial Intelligence Governance (Follow-Up)*, Report 2025-F-17, August 27, 2026, New York City Office of Technology and Innovation. Objective quoted from p. 4; other quotations from pp. 5–7. The report cites its authority as Article V, Section 1 of the State Constitution and Article III of the General Municipal Law (p. 3) and does not state that it was conducted under GAGAS. <https://www.osc.ny.gov/files/state-agencies/audits/pdf/sga-2026-25f17.pdf> and <https://www.osc.ny.gov/state-agencies/audits/2026/08/27/artificial-intelligence-governance-follow> — read 2026-09-27.

17. Madeleine Clare Elish, "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction," *Engaging Science, Technology, and Society* 5, 2019, 40–60, which describes how "responsibility for an action may be misattributed to a human actor who had limited control over the behavior of an automated or autonomous system." <https://estsjournal.org/index.php/ests/article/view/260>; abstract re-read via DOAJ (<https://doaj.org/article/97ff6743ea7a44a5ade2a04fd2c57a3c>), 2026-09-27. Applied here by analogy to the audit function.