AI Governance and the Architecture of Practice

James Maconochie

Technology Leader | BS Civil Engineering, Imperial College London ’93 | MS Civil Engineering, MIT ’94

May 2026

Abstract

AI is being deployed into judgment-laden institutional roles: clinical, financial, legal, regulatory, and academic, faster than governance frameworks for those roles are adapting. The frameworks being built treat AI-mediated judgment as an artifact-quality problem, manageable through inspection of what the AI produces and what the human signs. This whitepaper argues that they are designed against the wrong failure mode.

The variable that determines whether AI-mediated judgment is wise is not whether AI is used but how, whether the practitioner uses it as scaffolding for System 2 deliberation, or as amplification for System 1 reflex. This variable is invisible at the artifact layer. Two practitioners producing structurally identical compliant outputs can be running opposite cognitive architectures, and no inspection regime applied to artifacts will surface the difference. The failure mode that current frameworks cannot see, cognitive debt accumulating inside the humans the governance layer trusts to exercise judgment, compounds silently beneath the layer being inspected.

The whitepaper specifies four design requirements that any AI governance framework must meet to keep human wisdom in authority over the machine: cultivate the conditions under which AHI practice develops in practitioners, require its exercise in the moment, preserve the institutional conditions that sustain it across careers, and surface signatures of its absence at layers where artifact-inspection cannot reach. The four are necessary. Together they form a single integrated specification, not a list of independent recommendations.

The framework interlocks with a complementary architecture: the liability regime: tort, fiduciary duty, professional discipline, regulatory enforcement, that holds practitioners and institutions accountable after the fact. With both architectures in place, liability becomes diagnostic of architecture, and architecture becomes diagnostic of practice.

The argument is offered in the diagnostic register of structural specification rather than the prescriptive register of policy. It does not propose particular legislation, regulations, or institutional structures. It tells those who do govern what governance must do.

1. Introduction

AI is being deployed into judgment-laden institutional roles: clinical, financial, legal, regulatory, and academic, at a pace that exceeds the rate at which governance frameworks for those roles are adapting. This observation is widely shared. The response in policy and standards discourse, strongest in the EU's regulatory framework, in a number of US states, in sector-specific federal oversight, and in voluntary industry standards, has been a substantial expansion of the governance apparatus surrounding AI: disclosure requirements, audit obligations, bias-testing protocols, appeal pathways, model risk management standards, and transparency mandates. Where the expansion is occurring, the work is serious, and the practitioners doing it are competent.

The expansion is also the fundamental argument of this whitepaper, designed against the wrong failure mode. The governance frameworks being built treat AI-mediated judgment as an artifact-quality problem, something that can be managed by inspecting what the AI produces and what the human signs. Inspection regimes can be built around artifacts. They cannot be built around how a practitioner engages with an AI output before signing it. That engagement is not in the artifact. It is in the practitioner, and it produces no signal artifact-level inspection can detect. The frameworks are being designed against the failure modes they can see, while the one they cannot see accumulates beneath them.

This whitepaper claims that AI governance must be designed differently. Not more permissively, not less rigorously, and not on a different schedule, differently in its underlying architecture.

The whitepaper offers a governance specification that addresses this structural limit. The specification names four design requirements that any governance framework must meet to keep human wisdom in authority over the machine: cultivate, require, preserve, and surface. Each requirement addresses a specific failure mode that artifact inspection cannot detect. Frameworks meeting all four stand a chance of working. Frameworks meeting fewer than four will fail in the specific modes addressed by the omitted requirement(s).

What this whitepaper does not claim is also worth highlighting. It does not propose specific legislation, regulations, or institutional structures; the prescriptive register is structural, not policy-technical. It does not argue against deploying AI in judgment-laden roles; the argument concerns the conditions under which such deployment can be well governed, not whether it should occur. It does not survey the existing AI governance literature; the argument is presented as a position within a specific body of work on cognition and architecture, not as an integration of the field. And it does not claim to be the final word; the four design requirements are necessary, not sufficient.

2. Thesis

This whitepaper makes a structural argument about AI governance. The argument is prescriptive in its conclusions but diagnostic in its reasoning: it specifies what any

governance framework must do to succeed by first identifying the failure mode against which such frameworks must be designed.

The argument is grounded in a specific position. The author has spent the past year and a half developing a body of work on the architecture of human and artificial intelligence: the attention-experience feedback loop, the wisdom-knowledge gap, and the case for augmented human intelligence over artificial general intelligence. This whitepaper extends that work into governance. The position is one the author lives inside, and not always well.

2.1 The I-DEAS Story

In 1993, as a final-year civil engineering student at Imperial College London, the author chose a structural design project: an exhibition hall with a column-free interior under a long-span roof. The design tool of choice was I-DEAS, a finite element analysis package then considered cutting-edge, running on Silicon Graphics workstations that filled half a desk. The package allowed the structure to be modeled in full, member sizes specified, design loads applied, and the resulting stresses computed and visualized. The output was confident and articulate. The selected sections passed the relevant code checks. The deformed-shape projection looked plausible. The package said the structure would stand.

The night before the report was due, with no clear reason and no particular suspicion, the author did a back-of-the-napkin calculation of the stress at one of the two main arch supports. The number that came back was off, by a factor of ten, from what the package had reported. The discrepancy was traceable: the package had been working in newtons; the design loads, as entered, had been treated as kilograms. Off by a factor of ten in the wrong direction. Re-running the analysis with the correct unit assumption produced a deformed-shape projection that, in the package's own visualization, was flat as a pancake. The structure, as designed, would not have stood.

There was no time to redesign. The author wrote a preface to the report explaining what had been found, what it meant, and why the submitted design was not viable. The preface was thrown on the graders' mercy. The lesson the author took from it has shaped every encounter with a fluent technical tool since: the package was articulate, the output was internally consistent, the visualization was professional, and the answer was wrong. None of the surface signals the package produced contained the information that would have surfaced the error. Only an external check, made for no procedural reason and easily skipped, did.

The whitepaper's prescriptive argument rests on the proposition that this experience is general. The fluent tool produces confident output. The practitioner is not, by training or by credential, immune to ratifying it. Only architecture and individual practice, and the friction between them, can produce the gut-check that the tool itself cannot supply.

The thesis of this whitepaper is as follows.

AI governance that attends only to the institutional layer: audit, disclosure, appeal, bias-testing, will systematically fail to detect the failure mode current architecture cannot see:

cognitive debt accumulating inside the humans the governance layer trusts to exercise judgment. Because governance-by-inspection cannot distinguish AHI practice from its performance, institutional architecture must be designed to cultivate, require, and preserve individual AHI practice as its operative micro-foundation, and to surface signatures of its absence at layers where artifact-inspection cannot. Drawing on the loop architecture of The Wisdom Gap and on Hoze's structural-dissent principle, this paper sets out the design requirements for governance that keeps human wisdom in authority over the machine, not by inspecting outputs, but by architecting the conditions under which humans exercising judgment actually practice it.

The argument that follows is structured in four movements. Section 3 lays out the five structural claims on which the thesis depends. Each is necessary; together they establish what the variable is, why inspection cannot see it, and what the architecture must respond to. Section 4 demonstrates the structural argument concretely through worked examples in two domains. Section 5 specifies four design requirements that any AI governance framework must meet. Section 5I acknowledges what the framework cannot supply from inside its own walls and marks the boundary at which architectural specification gives way to the cultural, professional, and individual conditions on which any architecture rests.

The argument is offered as a reasoned position, the author believes, knowing the limits of his knowledge and cognitive capacity, and trusting that the reader will engage it with the same combination of confidence and humility with which it has been written.

3. Five Structural Claims

The thesis rests on five structural claims. They build on one another: Claim 1 establishes a categorical limit on governance-by-inspection; Claim 2 names what compounds beneath it; Claim 3 identifies the cognitive variable that distinguishes practice from performance; Claim 4 names the architectural principle that must replace reliance on individual virtue; Claim 5 specifies the composition relationship between the institutional and individual layers.

The claims are stated in their load-bearing form first; the reasoning that earns them follows.

3.1 Claim 1: The Invisibility Problem

Governance-by-inspection cannot distinguish AHI practice from its performance. Two humans producing identical compliant outputs can be running opposite cognitive architectures, one scaffolding System 2, one amplifying System 1, and the institutional layer has no instrument that sees the difference. Every design choice that follows rests on accepting this limit.

The limit is not a gap that better instruments will close at the layer where governance must operate: the individual decision, the individual practitioner, the moment of judgment. At

that layer, the limit is categorical. The cognitive process that distinguishes practice from performance happens inside the practitioner, in real time, as the AI output is read and absorbed. The process produces no external signal distinct from its performance. Audit logs record the decision, not the cognition that produced it. Documentation captures what was concluded, not how. Either practitioner can write reasoning notes; the practitioner who has interrogated the output and the practitioner who has ratified it can produce notes that are externally indistinguishable, because the reasoning that justifies a decision is reconstructible after the fact, whether or not it actually occurred before it.

It is categorical because the variable that matters does not live in the artifact. It lives in the practitioner. By the time the artifact exists, the variable is gone.

3.2 Claim 2: Cognitive Debt Compounds Silently at the Institutional Scale

Cognitive debt is the compounding interest you pay on a position you committed to without pressure-testing, on an argument accepted because it was fluent, a recommendation followed because it was articulate, a conclusion reached because the draft arrived already polished. The term refers to a specific failure mode, not a general concern about AI. It identifies what is incurred when a practitioner ratifies an output rather than interrogating it: a position is committed to, the institutional record reflects the commitment, and the cognitive work that would have stress-tested the position before commitment did not occur. The debt is the gap between what was committed to and what would have been committed to had the work been done.

Cognitive debt builds on Lisanne Bainbridge's "Ironies of Automation" (1983), which named the central paradox: the more thoroughly a system automates routine work, the more critical the human's residual role becomes, and the less practiced the human is at exercising it, because the routine work that built the practice has been automated away.

Cognitive debt is not reducible to Bainbridge's irony. It names what compounds at the institutional and career scale when disengagement becomes the operating mode of credentialed practitioners exercising judgment under AI mediation. And the property load-bearing for this whitepaper, that the failure mode is invisible to artifact-inspection, is the addition of cognitive debt that makes a tradition that has, until now, treated automation effects as observable in the work, the worker, or both.

The debt has three properties that distinguish it from related concerns.

It is invisible at the moment of incurrence. The practitioner who has ratified rather than interrogated rarely registers the difference. The felt experience is that of having made a decision. The artifact records a decision. The decision was made. What is missing is the work that would have made the decision well, and the absence of that work produces no signal that the practitioner can detect in the moment.

It compounds. A single ratified decision is recoverable; the practitioner can revisit it, and the institution can correct it. A practice of ratification across many decisions, sustained over a career, is much harder to recover, and the institution has no reliable mechanism for

doing so. The cognitive engagement that develops through repeated interrogation of difficult outputs does not develop. The judgment the practitioner is supposed to be exercising is not the judgment they have. The institution continues to trust the credential. The credential is now the wrong instrument.

It surfaces only at the institutional scale. One practitioner accumulating cognitive debt is a personal limit, regrettable but bounded. A cohort of practitioners doing so produces a cohort of senior practitioners who have never developed the practice. A discipline that does so produces a discipline whose collective critical capacity has been hollowed out. By the time the surfacing is unambiguous, the institution that would need to diagnose it is composed of the practitioners whose debt has rendered the diagnosis unavailable.

3.3 Claim 3: The System 1 / System 2 Distinction is the Operative Variable

The variable that determines whether AI-mediated judgment is wise is not whether AI is used but how, whether the practitioner uses the AI as scaffolding for System 2 or as amplification for System 1. This distinction is the operative variable that governance must architect around. The earlier essay AHI From the Inside developed it at length, drawing on Kahneman's two-system framework and on a specific case in which the author noticed, in real time, the pull toward a fluent reactive response and chose instead to use the available AI tool as a foil that opened deliberation rather than as an amplifier that polished a System 1 conclusion. The argument made there is presupposed here.

The two uses are not symmetric in the practitioner's experience. AI-as-scaffolding requires effort. The practitioner has to read the output as a starting point, ask what is missing, identify the load-bearing assumptions, and engage the deliberation that the output's fluency would otherwise foreclose. AI-as-amplification requires no effort. The fluent output arrives already shaped, already reasoned, already plausible, and System 1 reports that the work is done. The path of least resistance, in nearly every workflow AI is currently being deployed into, is amplification.

This asymmetry matters because the institutional pressures the practitioner operates within are themselves asymmetric. Throughput rewards amplification; deliberation costs time. Output quality, evaluated at the artifact level, is comparable between the two. Claim 1 established that the evaluation layer cannot distinguish between them. Career progression, evaluated on volume and timeliness of output, rewards amplification; the practitioner who insists on slow deliberation on every decision will be outcompeted, in any given quarter, by the practitioner who lets the AI do the deliberation. Without architectural counter-pressure, the asymmetry resolves toward amplification by default.

The whitepaper's prescriptive program rests on this claim. The four design requirements of Section 5 are responses to the asymmetry, not generic supports for "AI literacy" or "human oversight," but rather specific architectural counter-pressures designed to make scaffolding use possible against the systemic gradient that pulls toward amplification.

3.4 Claim 4: Structural Dissent as Architecture

Epistemic humility cannot be relied upon as a virtue in the practitioners whom institutions trust to exercise judgment. It must be architected as a structural feature of the institutional design itself, with dissent built into the workflow rather than left to individual courage. The reliance-on-virtue assumption is one of the most reliably failing assumptions in institutional governance: that practitioners who are credentialed, trained, and professional will, when the moment requires it, exercise the epistemic humility their role demands. Under throughput pressure, deadline pressure, and the everyday gravity of the path of least resistance, the assumption fails. It does not fail because practitioners are dishonest or unserious. It fails because virtue does not govern behavior under sustained pressure. Architecture does.

David Hoze, in his April 2026 essay extending the argument of The Wisdom Gap, named this principle directly: epistemic humility in beings of pure intellect, and by extension in any system whose default mode is fluent assertion, must be architected as a structural feature, not assumed as a virtue. Hoze drew on three thousand years of Jewish legal reasoning about how to govern intellectually capable agents whose capacity for self-correction cannot be presumed. The principle generalizes. Any system, human or artificial, whose surface output exceeds its internal mechanism for testing that output, requires structural dissent to function. The dissent is not a virtue the system might rise to. It is a feature that the architecture must build in.

The translation from system to practitioner requires care. System-level structural dissent, the adversarial process in courts, judicial review of agency action, and the Red Team in operations can be wholly external: the dissent happens between parties, none of whom needs to interrogate themselves. Individual-level structural dissent has no such option. There is no second practitioner in the workflow to serve as the adversary; the practitioner is the only cognitive agent positioned to dissent, and dissent must occur there or not at all.

This means the architecture cannot do the cognitive work for the practitioner. It can only make the failure mode harder to enact than the scaffolded alternative. The job of structural dissent at the individual level is not to substitute for cognition but to surface the cognitive failure mode at the moment it would otherwise compound, by requiring of the practitioner something only practice can supply. The foreseeable counter-strategy, friction performed without cognition, the cosmetic modification problem this paper later names, is the failure mode against which the design specification of Section 5 is built.

Translated to the practitioner working with AI, this principle has a specific implication. The practitioner's disposition to interrogate AI output cannot be relied upon, even in practitioners who are committed to interrogation in principle. The institutional architecture must structure the workflow so that dissent is operationally required, not solicited as a virtue but demanded as a procedural condition for completing the work. The practitioner who would, left to themselves, eventually drift toward amplification under pressure is held in scaffolding-use by the structure of the work itself. The architecture does the work that virtue cannot reliably do.

What the architecture must do is stop relying on virtue and start engineering for dissent. Section 5's design requirements are the operational specification of this principle.

3.5 Claim 5: The Composition Argument

Institutional governance and individual practice are not independent layers that combine additively. They work like reinforced concrete: institutional architecture is enormously strong in compression, the weight of rules, sanctions, and standards it can carry, but almost useless in tension. Individual practice is the reinforcing steel that gives the structure its capacity to bear tensile load: ambiguity, error, situations the rules did not anticipate. Each is necessary. Neither is sufficient.

The implication is that the prescriptive program of this whitepaper cannot be evaluated by examining either layer alone. A governance framework that meets all four design requirements but operates within an institution where individual practice has eroded will produce compliant artifacts and accumulating cognitive debt; the architecture will be in place, and the practice will not, and the architecture cannot supply the practice from outside the practitioner. A practitioner who exercises rigorous AHI practice inside an institution whose architecture has eroded the conditions for it will be outcompeted, ignored, or eventually worn down, the practice will be present and the architecture will not, and the practice cannot sustain itself without architectural support indefinitely.

The two layers are inextricably interdependent. Any prescriptive program that treats them as separable will fail at the very seam where they are not.

4. Worked Examples

Claim 5 stated the composition argument abstractly: institutional architecture and individual practice are interdependent, and governance-by-inspection cannot see whether the practice exists. The argument is abstract. The examples that follow are not.

Each example presents two practitioners who receive structurally identical AI-generated outputs and produce structurally identical compliant work. In each case, the governance layer cannot distinguish between them. In each, one has practiced AHI, and the other has accumulated cognitive debt. The two examples foreground two facets of the failure mode the whitepaper identifies: audit-blindness in heavily regulated industries, and the compounding of cognitive debt over institutional time.

The reader is not asked to find these examples surprising. The reader is asked to find them recognizable.

4.1 The Risk Score

A loan officer at a regional bank is reviewing an application for a $340,000 mortgage. The bank's underwriting platform has run the application through its risk model and returned a

recommendation: approve, with a risk tier of B+ and a suggested rate twenty-five basis points above prime. The reasoning surfaces the load-bearing factors, debt-to-income ratio, credit history depth, employment tenure, and loan-to-value, each annotated with its contribution to the score. The recommendation is consistent with the bank's risk appetite, with fair lending guidance, and is documented, traceable, and defensible.

Two loan officers, working the same queue, receive structurally identical outputs.

The first uses the recommendation as a frame for her own review. She reads the annotated factors and notices that employment tenure is contributing positively, despite the applicant having changed industries twice in five years. She pulls the employment record, sees the most recent role is six months old, and asks whether the model is treating sector continuity as a feature. It is not. She flags it, calls the applicant to discuss income stability, and recommends approval at a slightly higher rate. Her reasoning is documented. The model's reasoning is documented. The two are different, and this is evident in her file.

The second loan officer reads the recommendation, finds the factors reasonable, and approves at the suggested rate. He documents his reasoning by referencing the model's annotated factors. The file is complete. The decision is defensible. It will pass any audit that the bank's compliance function runs against it. It will pass any examination the regulator runs. It will pass fair lending review.

The governance scaffolding around consumer credit is among the most developed in any industry: ECOA, Regulation B, the Fair Credit Reporting Act, CFPB adverse action requirements, the disparate impact framework, and model risk management under SR 11-7. Every one of these instruments examines the output, the documentation, and the model. None examines the practitioner's cognitive engagement with the output.

The audit cannot see the variable that matters. The first and second loan officers produce files that are, to the auditor, indistinguishable. Both reference the model. Both document their reasoning. Both arrive at decisions consistent with the model's recommendation. The first has practiced underwriting. The second has performed it. The disparate impact framework, the most sophisticated lens any regulator brings to credit decisions, will register no difference between them. It is looking at the wrong layer.

When cognitive debt compounds at this layer, it does so within an institution that believes its governance has it covered. That belief is what makes the debt invisible. The 2008 mortgage crisis is the canonical illustration at scale, and the analogy is worth being precise about. The substance of the debt then was different: incentive corruption, no-doc loans, ratings agencies marking AAA on what was not, and the failure was knowing rather than blind. But the structural shape is what matters here. Risk had accumulated below the layer that the governance apparatus was inspecting. The disclosures, the ratings, the capital ratios, and the stress tests of the era all examined the artifacts the institutions produced and not the practices that produced them. When the debt surfaced, it did so everywhere at once because the inspection layer had never been able to see it accumulate. Cognitive debt in AI-mediated underwriting will not look like 2008 in its substance. It will look like

2008 in its shape: invisible until it isn't, distributed across an industry that believed its governance was sufficient, and concentrated, when it surfaces, in the populations the governance was supposed to protect.

4.2 The Manuscript

A reviewer for a mid-tier journal in computational biology has been assigned a manuscript: a methods paper proposing a new algorithm for protein structure prediction, benchmarked against three established approaches. The journal's review platform now offers an AI-assisted review tool. The tool ingests the manuscript and supplementary materials. It produces a structured review draft: an assessment of novelty, an evaluation of the benchmarking, questions for the authors, and a recommendation for major revisions. The draft is articulate, the questions reasonable, the recommendation defensible.

Two reviewers, working their respective queues, receive structurally identical drafts.

The first reviewer reads the draft as a starting point. She works through the methods herself, runs through the benchmarking choices, and notices that the comparison set excludes a recent approach from a competing lab that would be a natural inclusion. She adds a question about that omission. She also notices that one of the AI-generated questions misreads a notation convention in the field, and removes it. Her submitted review reflects her own engagement with the manuscript, scaffolded by the tool but not produced by it.

The second reviewer reads the draft, agrees with its assessment, edits two questions for tone, and submits it. The review is competent. The authors will receive useful feedback. The editor will receive a timely review. Nothing visible has gone wrong.

This is the example where the failure mode is hardest to instrument because no individual review fails. The marginal difference between the two reviews, in any single case, is small enough to dismiss.

The compounding is the problem.

The second reviewer reviews twelve manuscripts a year. Across a decade, that is one hundred and twenty manuscripts whose reviews were ratified rather than produced. His own engagement with the methodological frontier of the field, the kind that comes from working through other people's methods rigorously, repeatedly, over the years, does not develop. The skill peer review was supposed to cultivate in its reviewers atrophies. He continues to be assigned manuscripts because his prior reviews are on file. The credentialing stays current. The practice underneath it does not.

Across a field, across a decade, across thousands of reviewers, the compounding is structural rather than individual. The literature still gets reviewed. The reviews still arrive on time. The decisions still get made. But the institution of peer review, the slow, accumulative process by which a field develops its critical capacity through the act of reviewing itself, is hollowing out. No single review is a failure. The hollowing is the failure.

And no governance instrument operating at the unit of the individual review can see it, because the failure does not exist at that level. It exists at the unit of the reviewer's practice over time, and at the unit of the field's collective critical capacity over decades.

Two examples, two foregrounded patterns, one structural problem.

The two cases above demonstrate audit-blindness in a heavily regulated industry and the compounding of cognitive debt across institutional time. They do not exhaust the failure modes named in the whitepaper. The felt-fluency pull, the way a well-articulated AI output bypasses scrutiny by feeling complete, operates wherever practitioners receive synthesized recommendations under time pressure. The cosmetic modification problem, the engagement that looks like interrogation but is not, operates wherever signed work product is reviewed for fit and tone. The reader can supply the domain. The pattern is the same.

These are not four distinct failure modes. They are four faces of the same one. The governance layer is examining an artifact and cannot see the practice that produced it. The institution proceeds on the assumption that its inspection apparatus is sufficient, and the cognitive debt compounds beneath it.

If governance-by-inspection cannot see practice, the question Section 5 must answer is: what can?

5. Design Requirements

If the governance layer cannot inspect practice, the architecture must do something different. It must build the practice in practitioners, structure the work so that its exercise is necessary, protect the conditions that sustain it, and detect its absence at a layer where absence becomes visible.

What follows are four design requirements that any AI governance framework must meet to keep human wisdom in authority over the machine. They are stated as specifications rather than prescriptions. The whitepaper does not propose particular legislation, standards, or institutional structures. It proposes what any legislation, standard, or institutional structure must do to succeed, and, equivalently, the failure modes any framework must be designed to withstand.

The four requirements are: cultivate, require, preserve, and surface.

5.1 Cultivate

The practice is not a default state of the credentialed. It is a learned skill, and the conditions under which it is learned are specific: extended engagement with hard cases under graduated supervision, the experience of being wrong in ways that have consequences, the accumulation of judgment through repeated cycles of decision and feedback. These conditions have historically been provided by the structure of

professional development, apprenticeship, residency, junior associate work, and the early-career years in which practitioners encounter high volumes of unfamiliar problems with the latitude to engage them and the supervision to learn from being wrong.

The pressure AI exerts on this structure is not incidental. AI is most fluent at exactly the work the developmental pipeline depended on: synthesizing records, drafting first-pass analysis, and generating routine outputs. The institutional logic of replacing junior practitioners with AI assistance is, in the short term, defensible; the work gets done, costs come down, and throughput rises. The institutional logic of preserving the developmental pipeline is harder to defend in any given quarter, because the pipeline produces no measurable output of its own. It produces practitioners.

What this means in practice depends on the domain. In clinical training, it might mean requiring teaching institutions to demonstrate the structured experience junior physicians receive working through difficult cases without AI scaffolding. In credit underwriting, it might mean requiring institutions to document the developmental track for junior officers, including a defined volume of unassisted cases. In peer review, it might mean requiring journals to maintain mentor-reviewer pathways through which junior reviewers develop critical capacity on a cohort of unassisted manuscripts. The institution must architect for development that AI assistance would otherwise obviate.

A governance framework that meets this requirement must be architected against the short-term institutional logic. It must require that practitioners encounter those conditions even when they are operationally inefficient, and especially when AI assistance would obviate them. The specification does not require that AI be excluded from junior work. The specification is that institutions must demonstrate, at the architectural level, how their practitioners will develop the practice that the institution will later trust them to exercise. Without this, the senior practitioners of the next decade will be the loan officer who passed every audit, the reviewer whose practice hollowed out, the practitioner pulled by the fluent draft, the practitioner who edited for tone, at scale, and without the structural conditions under which recovery would be possible.

5.2 Require

Cultivation produces practitioners capable of the practice. It does not ensure they exercise it. The pressures of throughput, deadlines, and routine create a constant incentive for the practitioner to revert from scaffolding for System 2 to amplification for System 1, to read the AI output as thinking already done and sign their name. A framework that cultivates but does not require produces capable practitioners quietly converted into ratifiers by the workflow they operate within.

The architectural specification is structural friction. The work must be designed so that completing it requires the practitioner to do something only the practice produces, articulating the gap between the AI output and their own assessment, identifying what the system did not consider, and recording the substantive divergence between the recommendation and the decision. The friction must be load-bearing, not ceremonial.

Checkbox attestations that the practitioner “reviewed” the output do not meet this requirement. They produce an artifact that documents review without requiring it.

What this means in practice depends on the domain. In clinical decision support, it might mean requiring physicians to record what they would have ordered absent the recommendation, before seeing it. In credit underwriting, it might mean requiring loan officers to surface at least one factor the model did not weight. In peer review, it might mean requiring reviewers to flag at least one issue the AI draft did not raise. The architecture must demand of the practitioner something that ratification cannot supply, and must accept incomplete or formulaic responses as the failure they are.

5.3 Preserve

Cultivation builds the practice; operational necessity forces its exercise in the moment; preservation ensures it remains exercisable across a career. The conditions the practice depends on, adequate time per decision, supervisory cultures that tolerate error as a learning surface, peer review structures that test reasoning rather than ratify output, and professional norms that treat slow deliberation as competence rather than inefficiency, are the first casualties of throughput optimization. AI assistance, deployed without architectural protection, accelerates the casualty rate.

The pressures are not malicious. They are systemic. Institutions face genuine demand for faster decisions, lower costs, and higher throughput. The practitioner who insists on adequate deliberation time on every decision will be outcompeted, in any given quarter, by the practitioner who lets the AI do the deliberation. The institution that preserves the supervisory culture in which junior practitioners can be productively wrong will, in any given budget cycle, be outcompeted by the institution that automates around wrongness. Without architectural protection, these pressures compound until the conditions for practice no longer exist anywhere in the institution.

What this means in practice depends on the domain. In clinical care, it might mean specifying minimum time-per-decision floors that cannot be eroded without regulatory scrutiny. In credit underwriting, it might mean capping per-officer decision volume so throughput pressure cannot, by itself, consume the conditions for considered review. In peer review, it might mean limiting the number of reviews any single reviewer accepts in a year, with a corresponding floor on time per review. The architecture must hold the conditions for deliberation against the pressures that would otherwise consume them.

The specification is that governance frameworks must identify the institutional conditions on which the practice depends in a given domain, name them as load-bearing, and protect them with the same architectural seriousness applied to other irreducible institutional commitments. This is not a soft requirement to be balanced against efficiency. It is a structural commitment, on the same footing as audit independence or fiduciary duty, that the conditions for human judgment will not be optimized away, because once they are, recovering them requires the very conditions whose absence is the problem.

5.4 Surface

Even with cultivation, operational necessity, and preservation in place, the institution needs an instrument that detects when practice is failing.

Verification cannot occur at the artifact layer. Two artifacts produced by opposite cognitive practices are externally indistinguishable, and no inspection regime applied to artifacts will surface the difference. The verification must occur at a layer where the difference becomes visible, i.e., a layer where individual instances aggregate. The practitioner-over-time layer is one such layer. The cohort-of-practitioners layer is another. The discipline-over-decades layer is a third.

The architectural specification is that governance frameworks must include mechanisms operating at these aggregate layers, with the analytical capacity to detect signatures of practice and its absence. This is technically demanding and politically uncomfortable. It requires longitudinal observation of practitioners, which professional cultures resist. It requires statistical sophistication that current audit functions rarely possess. It requires the institution to act on signals that any individual practitioner can defensibly contest.

What this means in practice depends on the domain. In clinical care, it might mean monitoring, at a cohort level, the rate at which physicians' differentials align with AI recommendations, with thresholds below which the convergence no longer plausibly reflects independent reasoning. In credit underwriting, it might mean tracking how decision divergence from the model is distributed across officers and identifying clusters where the divergence is small enough to warrant review. In peer review, it might mean aggregating the rate at which reviewer questions are absent from the AI draft, across many reviews and many reviewers. The architecture must operate at the layer where practice and ratification become visibly distinct, without claiming the metric proves the absence of practice in any individual case.

But the alternative is the failure mode that the whitepaper has named. Without surface, the first three cannot be verified. Without verification, the framework is governance by aspiration. And governance by aspiration is the failure mode this whitepaper exists to identify.

Four requirements, three working for the practice and one working on it.

Cultivate builds the practice. Require forces its exercise. Preserve protects the conditions that sustain it across careers. Surface verifies that the first three are doing their work, because cultivation can fail, requirements can be reduced to ceremony, and preservation can erode under pressure. Without it, the institution cannot know whether its governance is functioning or merely performing.

Together, the four requirements specify what any AI governance framework must do to keep human wisdom in authority over the machine. They do not specify how. The how is domain-bound, institution-bound, and culture-bound. It is the work of the legislators,

regulators, and institutional designers who will translate these requirements into the particulars of medicine, finance, law, planning, science, and the domains beyond.

What this whitepaper claims is that no translation that omits any of the four requirements will succeed, and that the failure of any framework to address all four is diagnostic, not of imperfection, but of structural inadequacy. The four are necessary. They are not, by themselves, sufficient. What they rest on, and what no framework can supply from inside its own walls, is the subject of the next section.

6. What the Framework Rests On

6.1 Movement 1: Composition

The four design requirements of Section 5 compose into an architecture. Cultivate, require, preserve, and surface are not a list of independent recommendations but a single integrated specification. Each requirement closes a failure mode that the others would otherwise leave open. Cultivation without requirement produces capable practitioners who quietly ratify. Requirements without cultivation produce practitioners who cannot meet the friction the architecture demands. Preservation without surface produces slow erosion that no instrument can detect until the institution has already lost the capacity it was protecting. Surface without the other three produces an instrument that detects the absence of practice with no mechanism to restore it.

Together, the four requirements specify what an AI governance framework must architect: practices that exist, are exercised, are sustained, and are visible enough to be verified. This is what Section 5 claimed governance must do to keep human wisdom in authority over the machine.

It is not what governance must do to ensure that wisdom is exercised. The framework can be architected for the conditions under which practice is possible. It cannot architect the practice itself into existence in any individual practitioner. What it can do is necessary. What it cannot do, and what no architectural framework can do from inside its own walls, is the subject of what follows.

6.2 Movement 2: What the Framework Cannot Supply

The framework cannot supply the practitioner. It can architect the conditions under which practice develops; it cannot guarantee that any particular practitioner will develop it. It can structure work to require the exercise of practice; it cannot ensure that the exercise is genuine rather than performed. It can preserve the institutional conditions on which practice depends; it cannot legislate the cultural respect for slow deliberation that makes those conditions worth preserving in the first place. It can surface signatures of practice and its absence; it cannot generate the institutional will to act on what it surfaces.

Beyond the practitioner, the framework cannot supply the regulatory regime that creates it. It specifies what governance must do; it cannot generate legislators willing to legislate,

regulators with the authority and capacity to enforce, or institutions willing to accept the legitimacy of both. The nuclear precedent took decades of national legislation, international agreements, and inspection regimes invented before they could be applied. The four requirements assume the analogous regime for AI is being constructed. Without it, the specification has no body to translate it into law. At the US federal level, this presupposition is itself currently contested, a fact this whitepaper notes diagnostically rather than polemically, because a framework whose translation depends on a regulatory regime cannot be silent about whether that regime is being constructed or dismantled.

These are not gaps that better architecture would close. They are the boundary of what architecture can do. Architecture specifies conditions. It does not produce the human and cultural substrate that the conditions act upon.

That substrate has at least three components that the framework depends on but cannot supply. The first is professional cultures that treat the exercise of judgment as a craft to be developed across decades, not a credential to be issued at certification. The second is institutional leaders willing to bear the short-term costs of preserving the developmental pipeline against the constant pressure to optimize it away. The third, and the most consequential, is the individual practitioner's willingness to do the cognitive work the architecture makes possible, in moments where ratification would be easier and would pass every external test.

This third component opens a question that this whitepaper does not answer. The framework, as specified, addresses the institutional layer: what architectures must do to make practice possible, required, sustained, and visible. It does not address what these architectures produce in the practitioners who operate inside them over time, what sustained engagement with the practice develops in the human who exercises it across a career. That question, the practitioner-level output of a working architecture, distinct from its system-level output, is the subject of a forthcoming whitepaper. It is not the subject of this one. But it is worth marking the boundary, because the institutional argument made here is one half of what the larger framework is reaching for.

6.3 Movement 3: What the Framework Interlocks With

The four design requirements act on institutions prospectively: they specify the conditions under which AHI practice develops, is exercised, is sustained, and is verified. They do not specify what happens after the failure they were designed to prevent occurs. That work is done by a complementary architecture: the liability regime: tort, fiduciary duty, professional discipline, regulatory enforcement, that holds practitioners and institutions accountable for outcomes after the fact. The framework does not supply this architecture. It interlocks with it.

Cultivate interlocks with the standard of care. Where institutions can demonstrate the developmental architecture under which their practitioners' judgment was formed, the standard of care reflects something the institution actually built. Where they cannot, where the developmental pipeline has been optimized away, the institution's exposure

becomes structural. The harm is no longer attributable to any one practitioner's negligence. It is attributable to the institution that architected away the conditions under which non-negligent practice was possible.

Require interlocks with safe harbor. The practitioner who, prompted by the architecture, recorded what they would have ordered absent the AI recommendation, or surfaced a factor the model did not weight, has produced evidence of independent cognition. That evidence is differently positioned in litigation than its absence. It does not absolve the practitioner of poor judgment, but it distinguishes the practitioner who exercised judgment from the practitioner who ratified, in a way no after-the-fact reconstruction can. Required friction creates the artifact that makes liability allocation tractable.

Preserve interlocks with the reasonable-practitioner standard. The standard presupposes conditions under which a reasonable practitioner could exercise judgment, have adequate time, receive supervisory support, and have access to relevant information. An institution that has optimized those conditions away has made the reasonable-practitioner standard impossible to meet, and the exposure that follows is again structural: liability for the conditions created, not merely for decisions made within them.

Surface interlocks with discoverable evidence. The cohort-level signatures the requirement aims to detect, convergence patterns, decision-clustering, and atrophy of question-variation, are, when produced by the institution's own monitoring infrastructure, available in litigation. The institution that monitors its own practice has a different level of exposure than the institution that does not. The institution that monitors and finds nothing carries different exposure than the institution that monitors, finds something, and proceeds anyway. Surface produces the evidentiary trail that makes the prior three requirements adjudicable.

These four interlocks compose. The framework's architecture creates the institutional conditions under which liability can do its work; absent the framework, liability collapses into outcome-only judgment, with no instrument for distinguishing the institution that architected for practice from the institution that did not. With the framework in place, liability becomes diagnostic of architecture, and architecture becomes diagnostic of practice. The two architectures interlock, and neither is sufficient on its own.

6.4 Movement 4: What the Reader Can Do With This

What this whitepaper offers the reader is not a solution but a specification. Specifications do work. They tell the legislator what any law must require to succeed. They tell the regulator what any standard must measure to be diagnostic. They tell the institutional designer what any framework must build into avoid producing governance-by-aspiration. A specification does not govern. It tells those who do govern what governance must do.

The four requirements are the specifications outlined in this whitepaper. They are necessary. Any framework that omits one of them will fail in the failure mode that the omitted requirement was designed to address.

The reader who finds these requirements useful can do three things with them. The reader can use them as a diagnostic on existing governance frameworks, asking which of the four are addressed and which are missing. The reader can use them as a design brief for new frameworks, ensuring all four are architected from the outset rather than retrofitted. And the reader can use them as a vocabulary for naming failure modes that have, until now, been recognized in practice but not named in the governance literature.

What the specification cannot do is exercise itself. It can only be exercised, and the exercise is the work of those who govern. That work is now in front of them. This whitepaper has tried to make the work harder to avoid and easier to do well.

7. On Stability and Pressure

The architecture specified above must hold under pressure. A skeptical reader will ask why the four requirements, plus the liability interlock, do not, under real-world incentives, collapse into ceremonial compliance, metric gaming, and defensive over-documentation. The question is correct. The answer rests on the composition.

The four requirements are not independently stable. Surface alone invites Goodhart. Require alone invites box-ticking. Cultivate alone invites credentialism without practice. The liability interlock alone invites defensive over-documentation. The composition is what the architecture rests on: each requirement closes a failure mode that the others would leave open, and each pressure mode is bounded by a requirement designed to detect or punish it. A practitioner who games Surface still produces cognitive debt that Cultivate would have prevented from surfacing in misjudgment. An institution that ceremonializes Require still faces liability exposure when the ceremony fails to produce the engagement the standard of care requires. The architecture is adversarially stable, not because individual moves cannot defeat individual requirements, but because the moves that defeat one requirement are detected or punished by the others.

The architecture does not require a finished federal regulatory regime, a unified incentive model, or perfect signal reliability. It requires that some enforcement layer (regulators, professional bodies, insurers, or courts) interlock with at least some of the four requirements. Partial instantiation is a feature, not a flaw. A clinical setting where insurers enforce Cultivate, accreditation bodies enforce Require, and tort exposure enforces the liability interlock is a valid instantiation. So is a credit setting where regulators enforce Surface, and fiduciary duty enforces the rest. The architecture specifies the composition that the enforcement layer must assemble; it does not specify which institution enforces what.

Where enforcement layers are themselves degraded, with regulators captured, professional norms eroded, or courts politicized, the architecture has nothing to interlock with. The diagnostic register is provided so that failure modes remain named when the enforcement layers are under stress and addressable once the stress passes.

Three pressure modes will dominate any real instantiation: defensive over-documentation, metric gaming, and ceremonial compliance. Each is real, each is bounded by the composition, and each is the seam where the next iteration of the framework will do its work. The work this paper does not do is designing the operational details (the specific friction profiles, the specific signal-to-noise calibrations, the specific liability doctrines) that absorb each pressure mode in a given domain. That work is downstream of the specification.

The architecture does not promise stability under all pressures. It promises that the failure modes a stable architecture must absorb are now named and located. That is the precondition for designing for stability rather than presuming it.

8. Common Objections and Responses

A serious argument invites serious challenge. The following objections represent the strongest lines of resistance to the framework, addressed directly.

1. "The 'categorical' claim in Claim 1 overreaches."

Claim 1 has been narrowed to the operational layer: the individual decision, the individual practitioner, the moment of judgment. At that layer, the limit holds. Aggregate forensic methods may eventually surface population-level patterns across a practitioner's corpus; they cannot tell a regulator whether this practitioner exercised judgment on this case. The four requirements address what artifact inspection cannot do at the layer on which governance must operate. They do not depend on a categorical claim against all conceivable forensic methods. They depend on the operational claim, and it survives.

2. "Surface signatures invite gaming. Goodhart's Law applies."

It does, and the framework cannot pretend otherwise. What it can claim is that gameability is structurally bounded. Aggregate-layer signatures are harder to game than artifact-layer signatures: a defensible audit trail for a single decision requires only after-the-fact reconstruction; a defensible cohort-level distribution across many practitioners and many decisions requires coordination, the framework's other three requirements actively destabilize.

And Surface is not the binding constraint on the cognitive debt itself. A practitioner who games Surface produces the signature without the practice. The debt the practice would have prevented continues to accumulate, and it will surface eventually, in misjudgment under novel stress, in the failure of practitioners who never developed the engagement Cultivate was trying to build. Goodhart bites; it does not invalidate the architecture.

3. "Cognitive debt is one failure mode among many. Why privilege it?"

It is not privileged because it is the worst failure mode. It is privileged because it is the failure mode that the current architecture is structurally unable to detect. Bias, calibration error, misuse, hallucination: these are real, and the existing apparatus of disclosure, audit,

bias-testing, and appeal is built to address them. The apparatus has limits, but it has surface area: the failures it targets produce a signal at the artifact layer where the apparatus operates.

Cognitive debt produces no such signal. The whitepaper's claim is the narrower one: that any framework that addresses the visible failure modes without architecting for the invisible ones will fail at the layer it cannot see. That claim does not depend on cognitive debt being the worst failure mode. It depends only on cognitive debt being the failure mode, the apparatus being expanded to handle the others that it cannot reach.

4. "There's no incentive model. Why would actors actually build this?"

The objection lands. The whitepaper specifies what governance frameworks must do; it does not model the incentives of the regulators who would mandate, the institutions who would implement, or the practitioners who would comply. Behavioral economics has the vocabulary for that question. This paper does not deploy it.

The reason is that the paper is in the diagnostic register, not the prescriptive. It tells those who govern what governance must do. The prior question, whether the specification is right, has to be settled first. If the four requirements are wrong, no incentive model rescues them. If they are right, the incentive question becomes tractable in a way it is not currently, because actors gain what the diagnostic register provides: a target. The work of mapping incentives onto that target is real and necessary. It is not the work of this paper.

5. "Reinforced concrete displaced master masons. Common law supersedes prior precedents. Substrate change is sometimes net positive."

Sometimes it is. The argument is not that institutional conditions are always lost when superseded; rather, they are lost when the supersession fails to preserve the function the prior conditions served. Reinforced concrete absorbs the function master masons were exercising (structural load-bearing) and does so more reliably. Common law evolution absorbs the adjudication function: the new precedent decides what the old one would have decided. The substrate changes; the function is preserved.

In the architectures and use patterns this paper examines, AI mediation does not preserve the function of judgment. It produces outputs that look like the function's outputs without performing the function, the scaffolding/amplification distinction Claim 3 names. The objection succeeds against a paper claiming substrate change is categorically degrading. This paper claims something narrower: substrate change that produces the appearance of the function without the function is degrading, and current AI deployment patterns are doing the latter, not the former. The counterexamples are cases of the former. They do not address the claim.

6. "Surface cannot distinguish hollowing-out from AI-assisted excellence."

The objection lands. A cluster of practitioners with low decision-divergence from AI recommendations could be a cohort whose practice has hollowed out, or one that has

achieved high calibration with a model whose recommendations are correct. Surface alone cannot tie-break.

The architecture does not require Surface to tie-break. Surface is a flag, not a verdict. The verdict requires the composition: combining the Surface signal with evidence from Cultivate (whether the practitioners were trained in the engagement the institution claims they exercise), Require (whether the architectural friction was operational at the moments in question), and the consequences observed under novel stress (whether judgment was held when the AI recommendation was wrong). A signal that flags both hollowing-out and excellence is doing useful work, it identifies a cohort that the institution should look at directly. The risk to high-performing teams is real and bounded by the same composition that bounds the risk of false negatives. A framework that mistook the signal for the verdict would penalize excellence. A framework that uses the signal to trigger investigation and the composition to render the verdict does not.

7. "The standard-of-care interlock assumes a tort doctrine that does not yet exist."

It does. The claim that failure to architect for practitioner development becomes a breach of duty represents a non-trivial extension of how courts currently define negligence, moving exposure from the action layer (whether the practitioner deviated from peer behavior in this case) toward the institutional-architecture layer (whether the institution architected the conditions under which competent practice could develop).

The paper does not claim the doctrine is in place. It claims the doctrine is what the framework requires the liability layer to evolve toward, and that without that evolution, the prospective architecture is not enforceable. The interlocks named in Section 5I are forward-looking: they specify what each requirement asks of the liability regime, not what current case law already provides. Whether courts will move in this direction is a question for legal scholarship, the paper does not pretend to settle. What the paper does claim is the narrower, structural point: that without that movement, the architecture and the liability regime do not interlock, and the failure to interlock is itself a diagnostic of inadequate governance.

8. "The senior practitioners who would mentor are the cohort with the most cognitive debt."

The objection identifies the binding constraint on Cultivate's instantiation. Cultivate requires that institutions architect for practitioners to develop the practice, but the practitioners who would teach the practice are themselves the cohort whose practice may have hollowed out. The seed-corn problem named in The Wisdom Gap surfaces here as a transmission problem: the architecture for cultivation cannot assume the supervisory layer it depends on.

The framework's response is to name this as the binding constraint, not to dissolve it. Where the supervisory layer remains, Cultivate is implementable. Where it has hollowed out, Cultivate becomes a recovery problem, and the recovery requires identifying what survives of the prior practice. In these institutions, in which individuals, and architecting

around them while the next cohort develops. The architecture cannot recover from a generational break that is already complete. It can identify partial breaks, name what their completion would cost, and architect against that cost. Making the break visible at the layer where it can still be addressed is the framework's contribution. Pretending the break does not exist is not.

9. Intellectual Foundations and Prior Work

This paper does not claim to have originated the concern that human practice degrades under conditions of pervasive automation. The concern has a distinguished lineage in cognitive engineering and human factors research, and situating this argument within that lineage is both intellectually honest and practically important. The prior whitepapers in this series (AHI: First Principles, Maconochie 2026; and The Wisdom Gap, Maconochie 2026) draw on philosophy of mind and the architecture of biological intelligence. This paper draws on a more applied tradition: one that has studied the effects of automation in operational settings for four decades.

Bainbridge and the Ironies of Automation

Lisanne Bainbridge's "Ironies of Automation" (1983) is the primary intellectual ancestor of this paper. Her central observation, advanced in the domain of process control engineering, was that the more thoroughly a system automates routine work, the more critical the human's residual role becomes, and the less practiced the human is at exercising it, because the routine work that built the practice has been automated away. The irony is constitutive: the conditions that produce competent operators are the same conditions that automation removes. Two further ironies follow: errors that appear to be operator failures are often consequences of design choices that placed the operator in an impossible monitoring task; and as automation expands, the operator's role shifts from skilled engagement to vigilance, the cognitive disposition humans are worst at sustaining.

This paper extends Bainbridge in two directions. First, it names what compounds beneath her ironies at the institutional and career scale: cognitive debt, the accumulated cost of disengagement that surfaces in cohorts and disciplines rather than in individual operators on individual shifts. Second, it argues that the failure mode Bainbridge identified at the artifact-monitoring layer is being recapitulated at the judgment layer, where it is structurally more dangerous: the humans being asked to exercise vigilance over AI outputs are credentialed practitioners whose institutional authority rests on the assumption that they are exercising judgment rather than vigilance.

Automation Bias and Deskilling

Two adjacent literatures bear on the argument. The automation-bias literature, developed most fully by Parasuraman and Manzey (2010) and Skitka, Mosier, and Burdick (1999), documents the systematic tendency of human operators to over-trust automated outputs even when those outputs are demonstrably wrong. The effect is robust across domains,

training levels, and explicit warnings. The deskilling literature, descending from Braverman (1974) and extended through the human-computer interaction tradition, documents the long-term consequence: skills not exercised do not develop, and skills not developed cannot be exercised when the automation fails. This paper inherits both findings and situates them within the architecture: automation bias is what makes ratification feel like judgment in the moment; deskilling is what makes the institutional layer of practitioners progressively unable to do otherwise over the course of a career.

Clinical Decision Support and Alert Fatigue

The clinical decision support literature provides the most operationally documented case of the failure mode this paper identifies. Override rates in CDS systems routinely exceed 90% in deployed clinical settings; alert fatigue is the well-studied mechanism by which clinicians ratify or dismiss recommendations without engaging them. The literature is a working demonstration of cognitive debt at scale, in a domain where the artifact (the chart, the order, the signed note) gives no signal of whether the underlying judgment was exercised.

Aviation Automation

The aviation-automation tradition, most fully developed by Endsley (1995) on situation awareness and Sarter and colleagues on mode confusion and automation surprise, provides the longest case-study record of automation effects on expert practice. Its central finding is convergent with this paper's: as automation absorbs the cognitive work, situation awareness degrades, and the expert's capacity to intervene at the moment that matters atrophies. The aviation literature has produced the design-principle vocabulary that this paper extends to AI governance.

Dual-Process Cognition

The System 1 / System 2 distinction, as developed by Kahneman (2011), Stanovich, and Evans, is the operative cognitive framework Section 3 invokes. The earlier essay AHI From the Inside (Maconochie, 2026) developed the application of AI use at length; the framework is presupposed here.

Causal Reasoning and AI Limits

Judea Pearl's framework, developed across Causality (2000) and The Book of Why (with Mackenzie, 2018), provides the diagnostic vocabulary for what current AI architectures lack at the cognitive level. The Wisdom Gap (Maconochie, 2026) developed the application; this paper presupposes it and extends the institutional consequences.

Structural Dissent

David Hoze's argument that wisdom requires structural dissent, institutions designed so that disagreement is built into their architecture rather than dependent on individual

courage, provides the design principle Claim 4 inherits. The contribution of this paper is the practitioner-level translation: structural dissent at the system level does not automatically produce it at the individual level, and the architecture must be designed for both.

Attention and Experience

William James's observation that "my experience is what I agree to attend to," from The Principles of Psychology (1890), is the foundational claim the attention-experience feedback loop rests on. The Attention Crisis (Maconochie, 2025) developed the contemporary application; this paper extends it into the specific question of what governance must architect when attention is the upstream variable that determines whether judgment develops.

Prior Work by This Author

This paper is the fourth in a connected series. AHI: First Principles (Maconochie, 2026) provides the foundational argument for Augmented Human Intelligence as the correct alternative to the AGI scaling thesis and introduces the attention-experience feedback loop. The Wisdom Gap (Maconochie, 2026) develops the structural argument that current AI architectures are categorically ofincapable to producing wisdom and identifies the seed-corn problem that this paper inherits. The Attention Crisis (Maconochie, 2025) situates AHI within the broader challenge of infinite language production. AHI From the Inside (Maconochie, 2026) develops the System 1/System 2 distinction at the practitioner level. Earlier work includes Beyond Scale: Towards Biologically Inspired Modular Architectures (2025) and "LLMs, Synthetic Wisdom and Bias" (2024), the latter of which first introduced the concept of synthetic wisdom.

10. References

Bainbridge, L. (1983). "Ironies of Automation." *Automatica* 19(6), 775,779.

Braverman, H. (1974). *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*. Monthly Review Press.

Endsley, M. R. (1995). "Toward a Theory of Situation Awareness in Dynamic Systems." *Human Factors* 37(1), 32,64.

Evans, J. St. B. T. (2008). "Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition." *Annual Review of Psychology* 59, 255,278.

Hoze, D. (2026). "The Variable Somebody Already Measured." April 8. davidhoze.substack.com

James, W. (1890). *The Principles of Psychology*. Henry Holt & Co.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Maconochie, J. (2024). “LLMs, Synthetic Wisdom and Bias, and the Immediate Need for Independent Certification and Accreditation of Models.” LinkedIn, June 12.

Maconochie, J. (2025). *Beyond Scale: Towards Biologically Inspired Modular Architectures*. Architecture & Attention. jamesmaconochie.com

Maconochie, J. (2025). *The Attention Crisis: Language, Meaning, and the Architecture of Augmented Human Intelligence*. Architecture & Attention. jamesmaconochie.com

Maconochie, J. (2026). *AHI: First Principles*. Architecture & Attention. jamesmaconochie.com

Maconochie, J. (2026). “AHI From the Inside.” Architecture & Attention. jamesmaconochie.com

Maconochie, J. (2026). *The Wisdom Gap*. Architecture & Attention. jamesmaconochie.com

Parasuraman, R., and D. H. Manzey. (2010). “Complacency and Bias in Human Use of Automation: An Attentional Integration.” *Human Factors* 52(3), 381,410.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, J., and D. Mackenzie. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Sarter, N. B., D. D. Woods, and C. E. Billings. (1997). “Automation Surprises.” In *Handbook of Human Factors and Ergonomics*, 2nd ed., edited by G. Salvendy. Wiley.

Skitka, L. J., K. L. Mosier, and M. Burdick. (1999). “Does Automation Bias Decision-Making?” *International Journal of Human-Computer Studies* 51(5), 991,1006.

Stanovich, K. E. (2011). *Rationality and the Reflective Mind*. Oxford University Press.