Above-the-Loop Governance: A Statistical Architecture for Verifying and Allocating Decision Authority in Deployed AI Systems

Andrew M. Rochford

Independent Researcher, Sydney, NSW, Australia

ORCID: 0009-0005-8430-809X

andrewrochford@gmail.com

23rd June 2026

Abstract

Artificial intelligence systems increasingly make consequential decisions at a volume and velocity beyond human review and increasingly hold authority over those decisions rather than offering recommendations a human weighs. The dominant oversight model, the placement of a human in the decision flow, was designed for a different scale and is now constrained by documented cognitive limitations and by an economic structure whose cost grows linearly with deployment. This paper defines above-the-loop governance: a class of architecture in which qualified human professional judgement is not embedded in the per-decision flow but is held as a continuously measured benchmark against which an AI system's judgements are verified, using statistical methods established over more than half a century in clinical research, regulated manufacturing, and population-grade measurement. The architecture's defining economic property is that the human benchmark scales with statistical sufficiency rather than with AI volume, so the unit cost of governance falls as deployment grows under stable conditions. The paper extends the architecture in three respects: it anchors the human benchmark to externally established professional and regulatory standards, addressing the circularity of an organisation measuring itself; it introduces an authority layer that allocates decision authority between human and machine on measured evidence rather than by default; and it specifies the graded, accountable, measurement-verified response that must follow when divergence is detected. The paper situates the architecture explicitly within the scalable oversight literature of AI safety research and the post-deployment monitoring literature of regulated industries, states the assumptions and limits of each claim, offers a ten-criterion framework by which any system claiming the category should be evaluated.

Keywords: AI governance; human oversight; scalable oversight; inter-rater reliability; concordance; statistical process control; decision authority; agentic AI; regulatory assurance.

Acknowledgements

The author acknowledges the Centre for AI, Trust and Governance, Faculty of Arts and Social Sciences, The University of Sydney.

The author is particularly grateful to Dr Rob Nicholls, Senior Research Associate, Centre for AI, Trust and Governance, and Associate Professor Daniel Gozman, The University of Sydney Business School, for their review of and insightful comments on earlier drafts of this paper.

1 Introduction: two gaps in the governance of deployed AI

The governance challenge this paper addresses is structural rather than incidental. AI systems are increasingly making consequential decisions at a volume and velocity beyond the reach of case-by-case human review, and they are increasingly making those decisions on their own authority rather than as recommendations a human will weigh. The governance models developed for earlier generations of automation were not designed for either shift. The argument that follows identifies the missing architectural layer and specifies what a system delivering it must contain.

Two questions arise from this trajectory, and the second is, in current practice, answered by default rather than governed in real time. The first is whether the judgements an AI produces are accurate, safe, and reliable, and whether verification of that property keeps pace with the volume at which decisions are made. The second is prior to it: whether a given decision is one an AI may appropriately hold authority over at all, and under what continuous conditions that authority should stand or lapse. In current practice the second question is answered by configuration and incremental scope expansion, with no real-time mechanism to verify that the appointment remains appropriate or to withdraw it when the conditions that justified it no longer hold. This paper treats both questions as a single architectural problem.

1.1 The verification gap

Every organisation deploying AI into a consequential use case operates under the same constraint: the technology offers efficiencies a competitor will capture if it is not used, while the same technology produces outputs that, if wrong, expose the organisation to regulatory action, reputational harm, harm to those it serves, and litigation. This tension exists wherever AI participates in judgement flows that carry consequence, including clinical decision support and triage, lending and credit assessment, insurance underwriting and claims, fraud detection, legal triage, professional advice, regulatory compliance, and eligibility determination across public services. McKinsey’s State of AI 2025 survey, conducted across 1,993 respondents in 105 nations, reported that eighty-eight per cent of organisations now use AI in at least one business function, up from seventy-eight per cent a year earlier and twenty per cent in 2017.[1] Gartner forecasts that forty per cent of enterprise applications will embed task-specific AI agents by the end of 2026, up from under five per cent in 2025, and that at least fifteen per cent of day-to-day work decisions will be made autonomously by 2028, up from none in 2024.[2]

The deployment side is growing far faster than the side that verifies it. Stanford's AI Index 2026 records the same imbalance from the supply side: reporting on capability benchmarks is now near universal among leading developers, while reporting on responsible AI benchmarks remains uneven, and documented AI incidents rose to 362 in 2025 from 233 the year before.[3] Joint research by the United Kingdom’s Financial Conduct Authority and the Bank of England found that around three-quarters of UK financial-services firms now use AI, that fifty-five per cent of AI use cases involve some degree of automated decision-making, and that only two per cent are fully autonomous.[4] The two per cent figure is instructive in both directions: it

shows that genuine autonomy remains rare, and it shows that the prevailing safeguard for the other ninety-eight per cent is human involvement on individual decisions, a safeguard that does not scale. Removing that involvement without a defensible substitute is precisely the exposure organisations cannot accept; retaining it at scale is precisely the cost they cannot bear. That is the verification gap.

The gap is not a deficiency of intent. It is a mismatch in maturity. The infrastructure that governs the competence of qualified human professionals has developed across more than a century of regulatory and professional refinement; the infrastructure that governs AI decision-making has had perhaps five years. Applied to decisions of comparable or greater consequence, that mismatch is the structural problem this paper addresses.

1.2 The limits of human-in-the-loop, in their proper context

Human-in-the-loop oversight has been the right architecture for its era, and it remains the right architecture for many decisions. Placing a qualified human in the decision flow, with the authority to review, intervene, or override, has governed lower-volume and lower-stakes AI deployments with real rigour, and many organisations have implemented it conscientiously. The argument here is not that the principle is invalid; it is that scale has changed, and an architecture built for one scale encounters limits at another.

Those limits are documented. Article 14 of the European Union AI Act requires that high-risk systems be designed for effective human oversight and explicitly identifies automation bias as a risk overseers must be equipped to recognise.[5] The peer-reviewed literature, across several decades and multiple domains, defines automation bias as the tendency of operators to over-rely on automated output, such that the output becomes a heuristic replacement for vigilant information seeking and processing.[6] Experimental work has demonstrated automation bias and the conditions that intensify it in national-security decision settings,[7] and recent work in computational pathology documents both automation bias and anchoring, in which an AI recommendation presented before a human judgement shifts that judgement toward the machine, with anchoring intensifying under time pressure.[8]

Three further constraints arise specifically at scale. The first is the erosion of independent expertise: the ironies of automation identified four decades ago apply with renewed force, as humans who defer increasingly to AI on cases they no longer fully assess lose the practised capacity required to govern the system.[9] The second is the single-reviewer problem: most operational human-in-the-loop systems compare AI output to one reviewer, with no measurement of agreement between reviewers, leaving the standard against which the AI is judged itself unverified; the policy literature on human oversight of algorithmic systems finds that oversight requirements of this kind frequently fail to deliver the protection they are assumed to provide,[10] and empirical work on human-AI teams shows that the combination does not reliably outperform the better of its parts.[11] The third is the timing of detection: such systems typically detect problems after a wrong decision has propagated, rather than before divergence accumulates. Industry participants in the

Bank of England's AI forums have reached the same recognition in their own terms: members of the Bank's AI Consortium noted in February 2026 that maintaining a human in the loop may become increasingly strained as firms adopt agentic AI, and participants in the Bank's AI roundtables observed that the traditional model-risk emphasis on understanding how a model's inputs map to its outputs was not tenable or fully effective for increasingly complex AI models.[12]

> *Human-in-the-loop has been the right architecture for its era, and it remains necessary in many contexts. What has changed is scale, not the validity of the principle. The question is what additional architecture, operating alongside human oversight, makes governance complete.*

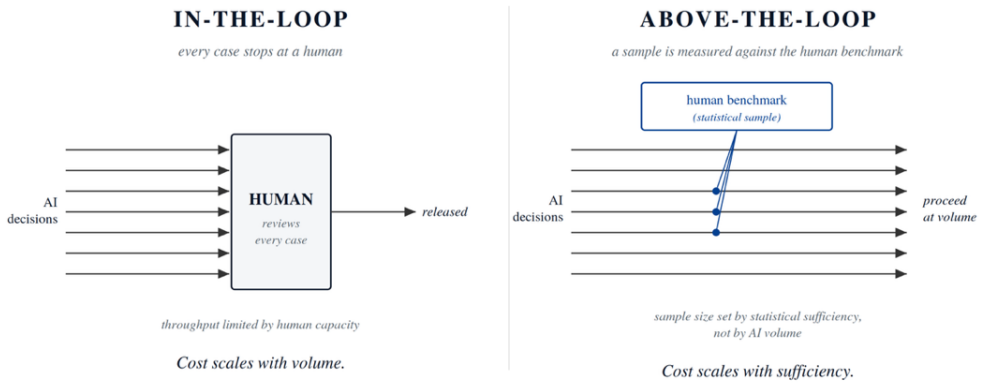

Figure 1. Two architectures, two cost structures. Under in-the-loop review every case is bottlenecked at a human, so the cost of governance scales with decision volume. Under above-the-loop governance the population proceeds at volume while a statistical sample is measured against the human benchmark, so the cost scales with statistical sufficiency rather than with volume.

1.3 The standard problem: what is the AI measured against?

A question the discussion of oversight tends to skip is what standard the AI is measured against. In the human-in-the-loop model the implicit standard is the judgement of the reviewing human, who serves as both safety net and benchmark. Two consequences follow. The standard becomes the judgement of one individual, whose own judgement is not verified against the broader professional standard; and anything inherent to the broader standard but not captured by that individual is not captured at all.

The standard against which AI in regulated contexts must be measured is not the judgement of any one person. It is the qualified human professional standard for that class of decision: the accumulated expertise of a profession, the institutional context in which decisions are made, and the human dimensions of consequential judgement.

That standard is not static; it evolves as practice evolves, and a system that measures AI against a frozen historical standard fails the moment the standard moves on.

1.4 From recommending to deciding: the authority gap

Alongside the change in scale, a change in kind is occurring. The earlier generation of consequential AI recommended and a human decided; the current generation increasingly decides, while a human, if present, supervises in aggregate or reviews after the fact. The scale of this shift is not speculative: Gartner forecasts that at least fifteen per cent of day-to-day work decisions will be made autonomously through agentic AI by 2028, up from none in 2024.[13] A system that recommends holds no authority; a system that decides holds authority, and any human role is then constructed around that fact rather than prior to it. An agent that resolves a decision autonomously has, by definition, been appointed to hold authority over it. That appointment is occurring continuously, through configuration choices and incremental scope expansion, rather than through any deliberate act of governance, and it is where the market is visibly failing. Gartner forecasts that more than forty per cent of agentic AI projects will be cancelled by the end of 2027, citing escalating costs, unclear business value, or inadequate risk controls among the causes.[13] Inadequate risk control, named by the analyst as a driver of cancellation, is the gap this architecture addresses.

The appetite to delegate authority to machines is rational, because machine-held authority is the configuration in which the value of agentic systems is realised; what is missing is a principled, auditable, continuous basis for deciding which decisions may be delegated, under what conditions, and with what real-time check that withdraws the authority when the conditions lapse. The existing oversight models do not provide it. Pre-deployment safety work establishes that a system is fit to operate but does not adjudicate, per class and continuously, whether it may hold authority; in-the-loop review reasserts human authority on every decision and forgoes the value of delegation; on-the-loop supervision observes aggregate behaviour after the fact and cannot make a forward determination. The appointment of authority is, at present, governed by nothing operating in real time.

2 The evaluation framework, stated first

A paper proposing a new governance category should invite scrutiny rather than defer it. This section states, at the outset, the framework by which any above-the-loop system should be judged; the remainder of the paper can then be read as a set of answers to it. The framework is offered to the field as a neutral instrument. Each criterion is derived from a documented failure mode of existing oversight, set out in Sections 1, 5, and 6, or from a stated regulatory expectation, rather than from the design of any particular system; it is not designed to advantage any provider, and it applies to every system claiming to deliver the category. The first five criteria concern whether a system is architecturally sound; the last five concern whether it can be operated, at scale and in production, defensibly. Both must hold: a system

satisfying the architecture but unable to operationalise it is of theoretical interest only, while a system that operates smoothly but lacks the architectural foundations produces output that is methodologically insufficient.

No.	Criterion	Where addressed
i	Independence of the human and AI systems, at the data layer	Section 4.1, 4.2
ii	An externally anchored, representative human standard	Section 4.3, 4.4
iii	Risk-weighted, information-valued sampling with design-weighted estimation	Section 5.2
iv	Continuous inter-rater reliability of the benchmark	Section 4.4, 5.2
v	Auditable, regulator-usable equivalence measurement	Section 5.3
vi	A defined, accountable response when drift is detected	Section 6
vii	Evidence-based allocation of decision authority	Section 7
viii	Honest treatment of the limits of human equivalence	Section 8
ix	Consistent, reliable operation at enterprise scale	Section 9
x	Deployment without new bureaucracy, with verifiable credentials	Section 9

Table 1. *A framework for evaluating above-the-loop systems, and where each criterion is addressed in this paper.*

The full framework, with the buyer guidance for applying each criterion, is restated in Section 10. Presenting it first is a deliberate signal: the architecture is offered to be tested, and the criteria by which it should be tested are published before the architecture is described.

3 The landscape of AI governance and the position of this work

Above-the-loop governance is one component of a broader landscape, and a proposal for a new architecture will be heard accurately only by an audience confident that the work already in place is understood. This section locates the principal components of that landscape, engages the two research literatures closest to this paper's claims, and positions above-the-loop governance as a complementary layer addressing a gap the others cannot fully close.

3.1 Governance across the AI lifecycle

Governance extends across an AI system's full lifecycle. Pre-deployment governance encompasses training-data curation, alignment, safety evaluation, red-teaming, and formal model validation; this is where most AI safety research has historically concentrated, and where it continues to concentrate.[14] Model-risk-management standards such as the United States Federal Reserve and Office of the Comptroller of the Currency's SR 11-7[15] and the United Kingdom Prudential Regulation Authority's SS1/23[16] span the lifecycle: they require pre-deployment validation and also mandate ongoing monitoring of models in use. Neither, however, specifies the statistical substrate by which that ongoing monitoring should be performed for AI systems, even where, as in SS1/23, the standard expressly addresses the use of artificial intelligence and machine learning in modelling. Calibration and re-training cycles continue after deployment on intervals defined in advance.

In-deployment oversight covers the moment-to-moment governance of the system in production, where the in-the-loop, on-the-loop, and out-of-the-loop models sit and where the limitations of Section 1 are most visible. Post-incident review completes the lifecycle. Each layer contributes work the others cannot replicate. The argument here is not that one layer replaces another, but that the in-deployment layer has lacked a continuous statistical verification component capable of operating at AI scale, and that above-the-loop governance is the design of that component.

3.2 The positions of human oversight, and the layer above them

Within in-deployment oversight, three positions for human judgement are established in the autonomy literature and in regulatory commentary, with their provenance in the analysis of autonomous systems[17] and meaningful human control.[18] Commentary on Article 14 of the EU AI Act maps the same positions onto regulated AI deployment.[19] In-the-loop places a qualified human in the per-decision flow; valuable at lower volume, it is constrained by the cognitive limits of Section 1 and by the single-reviewer problem. On-the-loop has humans supervise from outside the per-decision flow through dashboards and periodic sampling; its limitation is that aggregate supervision becomes informative only after problems are visible at the aggregate level. Out-of-the-loop has the AI operate autonomously within a defined scope; mature regulatory frameworks generally do not permit it for consequential decisions. The European Union's General Data Protection Regulation, for example, establishes a right not to be subject to a decision based solely on automated processing where that decision produces legal or similarly significant effects, a right the Court of Justice has read to reach even decisions on which a human formally signs off but in practice defers to the automated output.[20] The boundary is nonetheless being tested.

A note on terminology is owed before the term is used further. Phrases of the form "human above the loop" have circulated in the autonomy literature and in practitioner writing with varying senses, most prominently in recent management literature

describing the "agentic organisation," in which humans are positioned above the loop to oversee agents and workflows rather than to review individual decisions.[21] This paper defines the term architecturally and more demandingly: above-the-loop governance is not a seating position for a human but a measurement architecture, in which the qualified human standard is held above the AI's operation as a continuously verified benchmark, equivalence is measured statistically, and decision authority is allocated and withdrawn on that measured evidence. The ten criteria of Section 2 give the definition its content; a system that does not satisfy them is not delivering the category, whatever vocabulary it adopts.

Above-the-loop, so defined, is more than a fourth position. It is the meta-layer that governs which position each class of decision belongs in. The qualified human standard is held above the AI's operation as the benchmark; independent experts produce parallel judgements on representative samples, blinded to the AI's output; and a concordance engine measures equivalence continuously. The result of that measurement, combined with the authority layer of Section 7, determines whether a class should sit in-the-loop, remain under above-the-loop measurement with human authority, or be delegated to the AI under continuous above-the-loop verification.

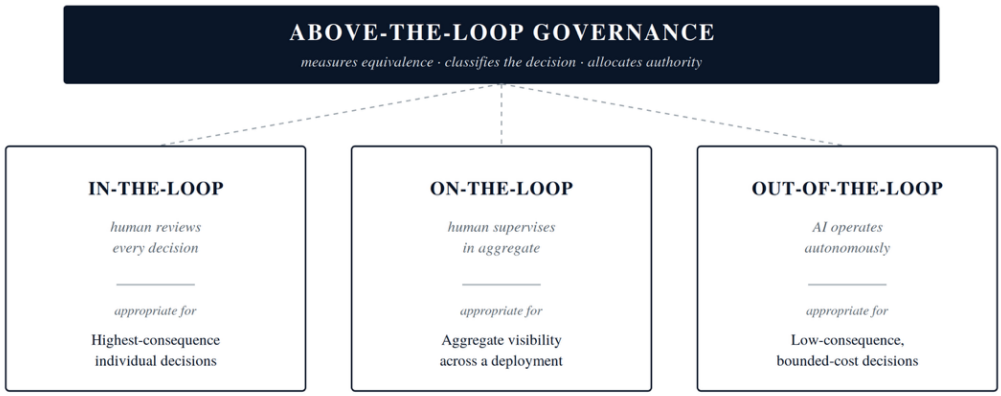

Figure 2. *Above-the-loop governance does not replace the established positions of oversight; it is the meta-layer that allocates each class of decision among them, on measured evidence.*

3.3 The scalable oversight literature, and the sense in which this paper uses the term

In AI safety research, scalable oversight names a research programme concerned with supervising AI systems whose outputs exceed the capacity of unaided human evaluation. The problem was named in that field a decade ago,[22] and the programme has produced a family of techniques: debate between AI systems before a judge,[23] iterated amplification,[24] recursive reward modelling,[25] the sandwiching paradigm for evaluating oversight protocols by positioning a model’s capability between non-expert and expert humans,[26] weak-to-strong generalisation, in which strong models are trained on labels from weaker supervisors,[27] and, most recently, quantitative

analysis of how oversight performance scales with the capability gap between overseer and overseen.[28] The adjacent AI control programme addresses deployment directly: it designs and red-teams control protocols, including trusted monitoring of an untrusted model's outputs, intended to remain safe even if the deployed model is intentionally subverting them,[29] extends those protocols to multi-step agentic settings,[30] and develops the safety cases by which a deployment under control measures could be justified.[31]

This paper uses the term in its older and broader sense: the statistical science of overseeing professional judgement at scale, developed over more than half a century in clinical research, regulated manufacturing, and population-grade measurement, comprising inter-rater reliability, designed sampling, and statistical process control. The two senses are complementary rather than competing, and the distinction is one of object and timing. The AI safety literature is principally concerned with producing reliable supervision signals for systems during training and evaluation, under the prospect that the system may come to exceed its overseer's capability; this paper is concerned with verifying, during operation, that a deployed system's decisions remain equivalent to an externally anchored human professional standard, and with governing the authority such a system holds. The safety field's own roadmaps identify precisely this layer as needed: scaling limited human supervision capacity across large deployments is acknowledged as an open requirement that training-time techniques do not themselves discharge.[32]

The nearest neighbour to this work is AI control, and the difference is the threat model. Control protocols are designed to be robust to intentional subversion: they assume the deployed model may be adversarial, may behave differently when it believes it is being tested, and may collude with monitors drawn from the same model family. Above-the-loop governance assumes the opposite and says so: it addresses the non-adversarial deployed system whose behaviour drifts with data, context, model updates, and the evolution of professional practice. That assumption is a stated scope condition of this architecture, not an oversight. Sampling-based verification is not robust to a system that strategically alters its behaviour on the cases it predicts will be sampled; where that threat model is live, control protocols are the appropriate complement, operating alongside the architecture described here rather than being replaced by it. Within its stated scope, which covers the overwhelming majority of present enterprise deployments, the contribution of this paper is to bring the established statistical science of human oversight to bear, as governance architecture, on the verification and authority questions that the AI-focused literatures leave open.

3.4 In-deployment monitoring and post-market surveillance: what exists, and what is missing

A second literature is closer still to deployment and must be engaged on its own terms. Machine learning operations has a mature body of work on monitoring models in production and detecting concept drift, covariate shift, and label shift.[33] In regulated medicine, post-market surveillance of AI is an active discipline:

algorithmovigilance has been proposed and developed by analogy to pharmacovigilance,[34] and the United States Food and Drug Administration's total product lifecycle approach, with Health Canada and the United Kingdom's MHRA, expects manufacturers of machine-learning-enabled devices to monitor deployed model performance and manage re-training risk, including through predetermined change control plans that specify in advance how post-deployment performance will be measured and modifications managed.[35] The European Union's AI Act imposes post-market monitoring obligations on providers of high-risk systems under Article 72 and use-and-monitoring obligations on deployers under Article 26, alongside the human oversight requirements of Article 14.[5] Management-system and risk standards complete the picture: ISO/IEC 42001 specifies the requirements for establishing, maintaining, and continually improving an AI management system,[36] ISO/IEC 23894 provides AI risk-management guidance,[37] and the NIST AI Risk Management Framework organises governance into the functions Govern, Map, Measure, and Manage, with continuous measurement a core requirement of the Measure function.[38]

The deepest precedent in this landscape deserves to be named directly, because the architecture of this paper transposes it almost component for component. Society authorises pharmaceuticals, products that can cause significant harm, for use in populations that will never individually be tested, and it does so with justified confidence because of the structure of the regime around them: authorisation is phased and earned on sampled evidence; manufactured batches are released on statistical testing of samples rather than inspection of every unit; surveillance after release is continuous and mandatory, with defined reporting obligations; and the authority to sell is conditional and revocable, withdrawn when the surveillance signal demands it. Phased authorisation corresponds to the parallel running and retrospective evidence of Sections 7 and 8; batch-release sampling to the continuous sampled verification of Section 5; pharmacovigilance to the drift detection and obligated response of Section 6; and conditional, revocable marketing authority to the authority layer of Section 7. The confidence the public places in an untested-on-them product is confidence in a sampling architecture anchored to an externally governed standard with mandatory post-market response. That settlement is what this paper constructs for a new class of decision-maker.

What this body of work establishes, individually and collectively, is the obligation to monitor and a toolkit for detecting change in a deployed model's inputs, outputs, and recorded performance. What none of it assembles is the architecture this paper specifies: a living human benchmark, composed of qualified experts, verified by inter-rater reliability, anchored to externally established professional standards, against which the deployed system's judgements are continuously and blindly compared; a graded, accountable, measurement-verified response obligation attached to detection; and an evidence-based mechanism for allocating and withdrawing decision authority itself. The instruments above define what an organisation must do; none supplies the continuous measurement substrate by which the doing is verified against the professional standard, and none governs the appointment of authority. Above-the-loop governance is offered as that substrate,

and the paper's claim of contribution is accordingly precise: not the novelty of any individual statistical instrument, which Section 5 is explicit in disclaiming, but the integration of those instruments with an externally anchored human benchmark and an authority layer into a single governance architecture for deployed AI.

Above-the-loop governance therefore does not displace pre-deployment safety work, calibration cycles, or human-oversight models. It is the continuous statistical layer that operates between calibration cycles and alongside whatever oversight model is in place, and where it detects emerging divergence, it provides the signal that a re-training cycle is required before divergence propagates into harm, making the broader system proactive rather than reactive.

4 The architecture: three systems in concordance

Above-the-loop governance is an architecture, not a position, and category claims that cannot be operationalised do not survive scrutiny. The architecture consists of three systems held in continuous statistical concordance: an established external standard, a composed human cohort that renders that standard live and scalable, and the AI. The measurement that matters is a single relationship, the AI against the composed human benchmark; the external standard and the composition of the cohort exist to make that benchmark the highest-quality, most independent, and most legitimate human standard obtainable.

4.1 Two systems, and the meaning of their independence

The measurement rests on two independent systems intersecting at a concordance engine. The human loop comprises qualified experts producing judgements on cases, independent of one another and of the AI; when they judge cases the AI has also judged, they are blinded to the AI's output at the data layer, not merely in the interface, so that anchoring and other contamination cannot affect their judgement. The AI loop is the AI system producing judgements within its deployed workflow; it does not see the human benchmark, does not pause for human review on individual decisions, and runs at whatever volume its deployment requires.

The independence claim requires precision, because a careless version of it is false and a careful version is the foundation of the method. The two systems are not statistically independent in the sense that the AI and the human profession share no common knowledge; plainly they do, since the AI was in general trained on artefacts the profession produced. The independence that the concordance measurement requires is narrower and exact: independence of the per-case judgement act. For any given case entering the sample, the human output is formed without sight of the AI's output and the AI's output is formed without sight of the human output, so that the two judgements on that case are not causally coupled. This is the condition that makes the agreement between them informative rather than circular, and it is an architectural property a buyer can inspect: that experts are blinded at the data layer, and that the AI does not receive the human benchmark as an input. Shared prior

knowledge does not defeat the comparison; coupling at the moment of judgement would, and the architecture is built to prevent it.

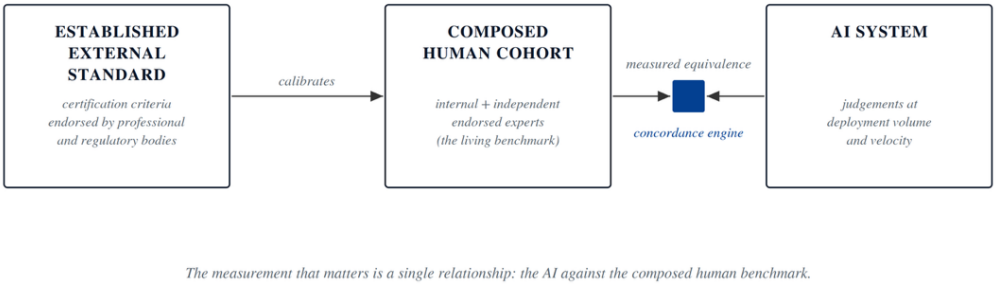

Figure 3. The three-system architecture. The external standard calibrates the human cohort; the cohort is the living benchmark; the AI is measured against it by the concordance engine.

4.2 The concordance engine

The engine performs three functions. It samples the AI loop by stratified, risk-weighted probability sampling rather than by simple random sampling, stratified by case class and weighted by consequence and information value, with sample sizes set by the statistical power required to detect material divergence.[39] It routes sampled cases to the human system for independent, blinded benchmark judgements, measuring inter-rater reliability at the case level where multiple experts judge the same case. And it measures equivalence between the human benchmark and the AI's output using established methods from clinical research and quality science.

The statistic is selected by the measurement level of the decision class, and the selection rule is fixed and stated: Cohen's kappa for categorical judgements,[40] weighted kappa for ordinal judgements, and the intraclass correlation coefficient for continuous judgements, with Krippendorff's alpha computed across the same cases as a measurement-level-general cross-check.[41] Bootstrap confidence intervals are computed on every update as descriptive summaries, decision-grade uncertainty is carried by the anytime-valid statistics described in the methodological note below, and values are banded against the widely used interpretation framework of Landis and Koch, with the qualification that such bands are conventions rather than derivations and that alternative conventions exist in the literature.[42] The concordance coefficient for a decision class is the single kappa-family value produced by this rule for that class. It exists per decision class; no weighted composite across classes is ever computed, and where a structured output is assessed field by field or dimension by dimension, any within-decision aggregation exists for reporting convenience only, while drift detection and the authority allocation of Section 7 operate on the per-dimension values.

The output is therefore a single, interpretable, statistically defensible value per decision class, expressing how equivalent the AI's judgement is to the qualified human standard on the sampled cases, with explicit confidence bounds reflecting the

volume and quality of the evidence on which it rests. Measurement begins from the first paired case, and the precision of the estimate improves as cases accumulate; below the volume required for confident interpretation the system reports a provisional state rather than a precise but unsupported value, and transitions to a settled state once the evidence supports it. The deploying organisation is never operating without verification, and the system never asserts precision the evidence does not support.

ON MULTIPLICITY, MATERIALITY, PREVALENCE, AND SEQUENTIAL INFERENCE

Four methodological cautions are intrinsic to operating such a system at scale, and an honest architecture states how it manages them.

First, multiplicity: continuous testing across many decision classes and on two detection surfaces invites false positives if each test is read in isolation. Thresholds are therefore calibrated for false-discovery control across the family of tests, not per test, and detection events are interpreted against that correction rather than against a naive per-test significance level.

Second, effect size versus significance: at high volume, trivially small divergences become statistically detectable. The architecture therefore distinguishes a detectable change from a material one, defining minimum effect sizes below which a divergence, however significant, does not trigger an obligated response; significance and effect size are always reported together.

Third, prevalence: kappa-family statistics are sensitive to the prevalence of outcomes, and the paradox of high observed agreement with low kappa under skewed prevalence is well documented.[43] Many of the decision classes this architecture will govern are heavily skewed, and a kappa time series can move because prevalence moved rather than because the system's behaviour changed.

The architecture therefore reports raw observed agreement and the proportions of positive and negative agreement alongside the chance-corrected coefficient, computes prevalence-robust companion statistics where skew is material,[44] and monitors prevalence as its own time series so that prevalence movement is identified as such rather than misread as concordance drift. Fourth, sequential inference: a system that monitors continuously must use statistics that remain valid under continuous inspection, because conventional confidence intervals examined at every update do not retain their nominal coverage when each look is treated as a fresh test.

The architecture therefore adopts the anytime-valid inference framework, descended from the sequential analysis founded by Wald, as its native inference engine: monitored values carry time-uniform confidence sequences, which hold their stated coverage at every moment of inspection simultaneously, and threshold judgements are expressed through e-values, whose evidence accumulates and combines across decision classes and can be acted upon at any time without inflating error.[45] The run-length-calibrated drift surfaces of Section 5.2 operate alongside this engine, and conventional bootstrap intervals are retained as descriptive summaries of the accumulated evidence, never as decision instruments.

4.3 Anchoring the human standard to an external benchmark

The most consequential question in the architecture concerns where the human standard comes from. If the benchmark were simply the deploying organisation's own qualified cohort, the standard would be set, measured, and audited by the same organisation. However qualified its staff, that is a form of self-assessment, and a standard an organisation defines for itself, measures itself against, and audits itself is not independently legitimate, no matter how rigorous each step.

The resolution the architecture specifies is to anchor the human standard to something external that already exists. The established external standard, the codified practice and certification criteria that professional bodies and regulators have already endorsed, is the legitimate but slow and unscalable baseline. The qualified human cohort is then the living, scalable instance of that standard, composed of two sources: the organisation's own internal experts, and independent experts nominated or endorsed by the relevant professional and regulatory bodies. The independent experts are blinded to both the AI and the organisation, so that neither institutional nor company-specific bias contaminates their judgement, and they enter through the same blinded data-layer flow as internal experts, judging cases in their native form.

The engagement model for the independent experts is specified, because independence that cannot survive scrutiny of who pays for it is not independence. Endorsement comes from the relevant professional or regulatory body, which nominates or accredits experts to be eligible for a panel register; where no body yet operates such a register for a domain, an independent institution of recognised standing acts as the endorsing intermediary, and the architecture reports the maturity of the endorsement honestly. Payment is structured against influence: experts are paid fixed per-case fees at rates published in advance, funded by the deploying organisation into a ring-fenced pool and administered by the platform operator as paying agent only, with blinded case allocation, no disclosure of client identity, no compensation linked in any way to the content of judgements, and caps on the share of any expert's panel income attributable to a single deployment. The fee is fixed before the judgement is formed and nothing about the judgement affects future fees, which is the same logic by which audit independence is protected. Composition carries a stated floor: for any decision class granted or under consideration for conditional machine authority under Section 7, independent external experts must contribute no less than one third of the benchmark judgements, and all certification calibration cases are externally sourced.

The external pool is itself calibrated against the established standard. The certification reference cases that professional and regulatory bodies already maintain to certify human practitioners, cases carrying an externally established correct judgement, are interwoven, blinded, into the external experts' caseload on a statistically chosen cadence, and the pool's concordance with those reference cases is measured by the same engine. High concordance confirms the pool is a faithful, certified instance of the external standard. The chain of legitimacy is thereby

completed: an established external standard, faithfully reproduced by a calibrated and independent human cohort, against which the AI is continuously measured.

The honest boundary is that the external anchor is only as strong as the existence of certification reference cases for a given domain. Where a profession certifies its practitioners against an established case bank, the cohort can be calibrated to it and the anchor is firm; where a profession maintains no such bank, the independent experts still cure institutional bias by virtue of their independence but cannot be certification-calibrated, and the architecture reports the weaker anchor honestly rather than claiming the stronger one.

WHAT THIS CONSTRUCTION DOES, AND DOES NOT, RESOLVE

It resolves self-assessment circularity and institutional bias. The benchmark is no longer set by the deploying organisation alone; independent, externally endorsed experts, calibrated against certification standards and engaged under the influence-resistant model above, mean the organisation no longer marks its own work, and divergence between the internal and external cohorts renders organisation-level bias visible and measurable.

It does not resolve profession-wide bias. If an entire profession shares a systematic error, independent experts drawn from that profession share it, and certification reference cases authored by that profession may encode it. Catching profession-wide error requires anchoring to realised outcomes where they are recoverable, addressed in Section 8. The architecture states this limit rather than obscuring it.

4.4 A representative, verified, adaptive benchmark

Because the benchmark is produced continuously by qualified professionals operating in their institutional context and anchored to an external standard, it is both verified and adaptive. It is verified because inter-rater reliability is measured wherever more than one expert judges a case, using Fleiss' kappa and complementary statistics, so the benchmark is protected from individual judgement variance and a record is kept of which case classes carry higher inherent variance among qualified experts.[41] It is adaptive because as professional practice evolves the cohort's judgements move with it, and the AI is held to the current standard rather than a historical one. Where the cohort's evolving practice begins to diverge from the codified external standard, that divergence is itself measurable and becomes evidence the organisation can return to the professional body that the codified standard should be revisited, making the architecture a mechanism by which the external standard is kept current rather than a frozen reference that ages.

5 Methodological foundations

The defensibility of the architecture rests not on novel statistical theory but on the application of established methods, refined over decades and accepted across the regulated decision-making landscape, to the problem of AI verification. This section sets out the principal foundations, the literature on which each rest, and the assumptions each carries.

5.1 Population sampling, non-stationarity, and why the human system does not scale with volume

A properly drawn sample of statistically appropriate size supports defensible conclusions about the population from which it is drawn. The principle underwrites clinical-trial design, manufacturing quality assurance, and national-scale measurement, and the methodology for determining sample size is itself well established: Cochran's formula gives the size required for a stated confidence level and margin of error,[39] stratified sampling ensures sub-populations of interest are represented, and statistical power analysis determines how much data is required to detect a meaningful difference. Applied here, the human system is a sample of qualified expert judgement, of statistically appropriate size, stratified by case class; the AI system is the population. As the AI system grows, the absolute number of cases the human system must judge for a given confidence level grows slowly and then plateaus, so the sampling ratio falls as confidence accumulates. A human system, a small fraction of the size of the AI system can govern it, because the relationship between sample and population permits it, exactly as a properly drawn audit sample supports defensible conclusions about a full ledger of transactions.

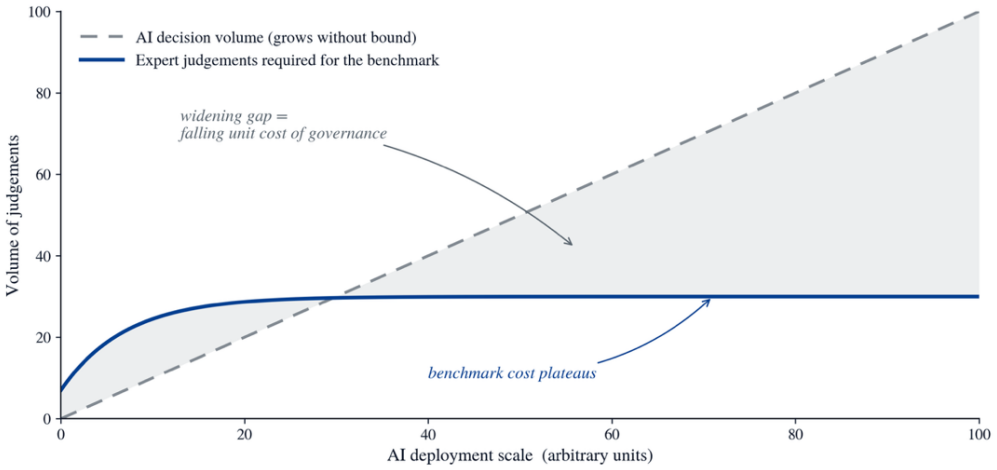

Figure 4. The scaling relationship, shown schematically. The expert judgements required for a defensible benchmark rise and then plateau, while AI volume grows without bound; the widening gap is the falling unit cost of governance. The curves are illustrative of the theoretical relationship, not measured output.

This relationship carries three assumptions that must be made explicit. The first is non-stationarity. Classical sampling assumes a stationary population, whereas the premise of Section 6 is that AI output distributions drift, which is to say the population is not stationary. The two are reconciled as follows. The plateau describes the sample required to characterise the population at a given time, under the prevailing distribution; it is not a claim that a fixed historical sample governs the system forever. The architecture maintains the benchmark as a rolling sample, continuously refreshed, and the drift-detection surfaces of Section 5.2 are precisely the mechanism that signals when the distribution has moved enough that the current sample must be re-weighted or enlarged. The sub-linear scaling claim is therefore a claim about the steady-state cost of maintaining a current benchmark under bounded drift, not a claim that non-stationarity can be ignored. Where drift is rapid and sustained, the required sampling rate rises for as long as the instability persists, and the architecture reports that the benchmark is being maintained under elevated cost rather than concealing it.

The second assumption is a bounded decision-class taxonomy. Sample size plateaus per decision class; total expert demand grows with the number of classes governed. The economics of Section 9.2 hold where the taxonomy grows slowly relative to volume, which is the observed pattern in operational deployments, and the architecture reports expert demand per class so that taxonomy growth is visible as a cost driver rather than absorbed silently into the scaling claim.

The third concerns detection latency, and it is the assumption a safety-minded reader should press hardest. Sampling sufficient for estimation is not automatically sufficient for timely detection: at a fixed absolute sampling rate, the number of decisions released between the onset of drift and its detection grows with deployment volume, so the unit cost of governance can fall while the exposure per drift event rises. The architecture therefore sets the sampling rate for each class against two requirements jointly: the precision required for the concordance estimate, and a maximum tolerated number of decisions at risk between drift onset and expected detection, set according to the consequence of the class. Sampling scales with exposure tolerance, not merely with estimation sufficiency, and unlike per-decision review, whose cost is structurally linear in volume, this remains a sub-linear and explicitly governed cost.

5.2 Inter-rater reliability, information-value sampling, and drift detection

Inter-rater reliability measures agreement between independent judges evaluating the same cases and is the method by which the integrity of the human benchmark is itself verified. Cohen's kappa, introduced in 1960, is the most widely used measure for categorical judgements, correcting for chance agreement;[40] the Landis and Koch framework provides widely used interpretation thresholds, with the caveats of Section 4.2;[42] Fleiss' kappa extends the approach to multiple raters, weighted variants account for the magnitude of ordinal disagreement, the intraclass correlation coefficient performs the analogous function for continuous judgements, and Krippendorff's alpha generalises across measurement levels. Sim and Wright set out

the use, interpretation, and sample-size requirements of the kappa statistic in reliability studies; the volume of paired cases required for a target confidence-interval width depends on the expected level of agreement and the marginal prevalence of outcomes, and the architecture computes that requirement per class rather than assuming a universal figure.[41]

Sampling is weighted by the information value of individual cases, drawing on three established literatures: Shannon's information theory, which formalises why low-confidence cases carry more information per unit;[46] the active-learning literature, which shows that prioritising uncertain and atypical cases accelerates calibration;[47] and regulatory guidance on risk-based monitoring, which establishes that monitoring intensity should be proportionate to consequence.[48] Each AI decision is scored on uncertainty, consequence, and novelty, and the score sets the probability that the case is routed to expert review. A sample weighted toward hard and consequential cases is, by construction, not representative of the population, and an estimate computed naively upon it would understate population concordance and shift whenever the routing policy shifted. The architecture therefore records the inclusion probability of every sampled case and computes population-level estimates with inverse-probability design weighting, in the manner of Horvitz and Thompson,[49] and routing policies are versioned so that estimates remain comparable across policy changes. Unweighted estimates on the information-valued sample are retained as a deliberately conservative view of performance on the hardest cases, and the two are never conflated.

Drift is detected by statistical process control, drawn from quality-control science. Cumulative-sum control charts, first described by Page in 1954,[50] accumulate evidence of deviation from a reference value and signal when it exceeds a calibrated threshold. The design is stated rather than gestured at: for each monitored series, a reference value is set from the in-control baseline established during calibration, the allowance parameter is chosen for the smallest material shift the class is required to detect, consistent with the materiality thresholds of Section 4.2, and the decision interval is calibrated to a target in-control average run length, so that the expected frequency of false alarms and the expected delay in detecting a material shift are both design quantities, set per class according to consequence, rather than emergent accidents. The severity bands of Section 6 are mapped from the magnitude and persistence of the signalled excursion under that design. The same anytime-valid discipline set out in Section 4.2 governs the monitored series themselves: wherever a value is watched rather than merely reported, time-uniform confidence sequences replace repeated fixed-sample intervals, so that validity is a property of the design rather than a correction applied to it.[45] The architecture applies this on two complementary surfaces: a concordance surface, monitoring the time series of the concordance coefficient on the calibrated subset of cases, and a population-behaviour surface, monitoring the distribution of the AI's outputs across the full set of decisions, including signals such as output and confidence distributions and the rate of human overrides in any in-the-loop channels. The two are complementary because a system can hold stable concordance on its sampled cases while drifting in its overall distribution, and the reverse. Bland-Altman analysis provides the standard

representation of agreement between two measurement methods on continuous outcomes.[51]

5.3 Auditable evidence of equivalence

The output is a continuous, statistical, auditable record. For any governed deployment, the record contains which cases were sampled, how, and with what inclusion probabilities; the human benchmark for each sampled case and its inter-rater reliability where multiple experts contributed; the AI's output on the same case; the equivalence measurement and its confidence interval given the evidence; and whether drift has been detected on either surface. In the language of regulated industries this is audit-grade evidence: the record that supervisory expectations of continuous monitoring increasingly demand, that risk officers need to defend deployments to boards and auditors, and that investigators need when a decision is challenged and the question is whether the AI was operating within the established standard at the time. For all the machinery beneath it, the record's external interface is deliberately small: for any decision class it reduces to a single, current, revocable attestation, that the class is operating within the externally anchored standard at a stated authority tier, with the full statistical trail standing behind that attestation for any party entitled to inspect it. Section 6.4 makes the value of this record concrete.

6 The response when divergence is detected

Detection without a defined response is only a better alarm. An architecture that detects divergence with precision but leaves the response undefined produces an excellent early-warning signal and no governance. This section specifies the response: what happens when drift is detected, who is accountable, within what time, and how the remediation is itself verified. Supervisory attention falls precisely here.

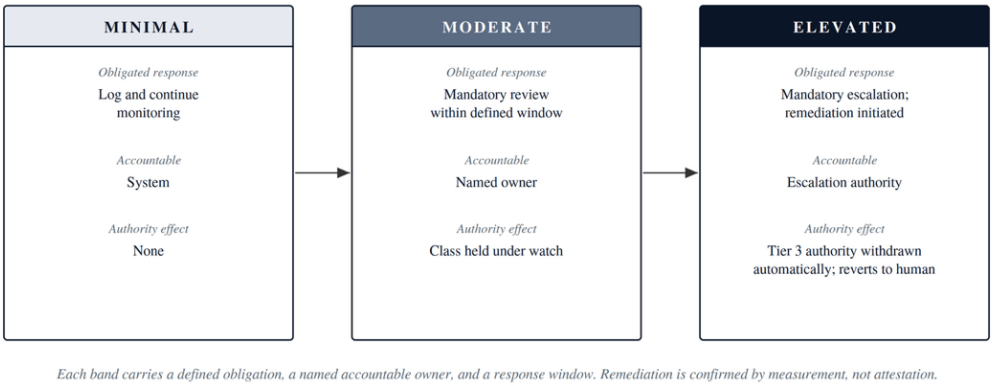

Figure 5. Each severity band carries a defined obligation, a named accountable owner, and a response window. Elevated drift triggers automatic withdrawal of conditional machine authority, and remediation is confirmed by measurement rather than attestation.

6.1 Graded response, tied to severity

Drift is not a binary alarm. The architecture grades its response to the severity of the detected drift, mapping a defined obligation, a named owner, and a maximum response window to each band, and using the same severity bands the detection engine produces under the run-length design of Section 5.2.

6.2 Named accountability, mandatory windows, and verified remediation

Each severity band is assigned to a defined organisational role rather than to diffuse responsibility. The accountable parties are the organisation's own governance roles; the architecture does not invent new institutional offices but records which existing role is accountable for each band, and records whether the response occurred within the window. This converts a stated process into a timestamped record that the process was followed on every occasion, which is the difference between asserting governance and proving it.

Critically, remediation is not marked resolved on the strength of the action taken. When drift triggers a response and a remedy is applied, the issue is closed only when the concordance engine confirms that equivalence has returned to the required band and held there for a defined confirmation period. Remediation is verified by the same measurement that detected the drift, not by attestation that something was done; the loop closes on evidence, producing the record a regulator expects: drift detected, response within window, remedy applied, equivalence confirmed restored, with the full statistical trail.

THE HONEST LIMIT OF THE RESPONSE ARCHITECTURE

The architecture guarantees that drift is detected that a response is obligated and timed, that it conditional machine authority is withdrawn where stakes are highest, and that resolution is confirmed by measurement. It does not guarantee that the remediation chosen is the correct one; selecting the right remedy remains the deploying organisation's professional responsibility. What the architecture guarantees is that the signal cannot be ignored and that resolution cannot be declared without the measurement confirming it.

6.3 Automatic withdrawal of authority

Where a class has been granted conditional machine authority and drift on that class reaches the most severe band, the architecture withdraws the machine's authority automatically and reverts decisions to the human, without waiting for human action to trigger the reversion. Machine authority was always conditional on continuous equivalence, so severe drift does not merely raise an alert on such a class; it lapses the condition on which the authority depended. The default on any uncertainty is that authority reverts to the human. This is the response architecture and the authority layer of Section 7 operating as one mechanism.

6.4 Accountability under challenge: a worked scenario

Consider an AI-assisted decision challenged eighteen months after it was made, in a court, before a regulator, or in a parliamentary inquiry. The question put to the organisation is whether the decision was produced within the established professional standard at the time. Under conventional oversight the organisation can produce, at best, the record of a single reviewer's sign-off, against a standard that was never itself verified. Under above-the-loop governance it can demonstrate something materially stronger: for the relevant class at the relevant time, the concordance coefficient and its confidence interval, evidencing statistical equivalence to the qualified human standard; the inter-rater reliability of that benchmark, evidencing that the standard was verified and not the view of one individual; the external anchor, evidencing calibration to the certification criteria of the relevant body rather than to the organisation's own definition; the authority tier the class occupied, evidencing that the decision was held by the party that should have held it; and the absence of unremedied drift, evidencing that the system was operating within standard at the time. That is the record of stewardship a challenged organisation needs, produced continuously, at scale, without slowing the AI. The weight such a record carries in any proceeding is, of course, for the court, tribunal, or supervisor to determine; the architecture's contribution is that the record exists, is statistical rather than testimonial, and was produced in real-time rather than reconstructed after the fact.

7 The authority layer: allocating decision authority on evidence

Equivalence measurement establishes whether the AI judges as the qualified human standard judges. It does not, on its own, establish whether the AI should hold authority over the decision. Those are different claims, and the market is currently answering the second by default. The authority layer governs the appointment of decision authority between human and machine, on evidence and continuously; it is the capability that turns measurement infrastructure into a governance mechanism for authority itself.

7.1 The keystone: measurement standard versus decision authority

The layer rests on a single distinction. The measurement standard is what the AI is measured against; in this architecture it is always anchored to the qualified human standard, and never to the AI's own output. This is what is fixed: the anchor is human, and it does not migrate to the machine as the machine gains authority. The standard so anchored is not frozen. As Section 1 noted, it evolves as professional practice evolves, and, in the decision classes where reliable realised outcomes are recoverable, those outcomes are the stronger anchor still, calibrating and where necessary correcting the human standard, as Section 8 sets out. What does not change is that the benchmark is external to the AI: the machine is never measured against itself.

The decision authority is the party whose judgement stands as the operative decision; this is not fixed but allocated per class on evidence. Humans always set the standard the AI is measured against. Who holds authority over the decision is a separate question. The AI is measured against the human standard in every case, including the cases where the AI holds authority, because that measurement is the continuous condition on which the authority depends. The standard does not recede when the AI holds authority; it becomes the live test the AI must keep passing to retain it.

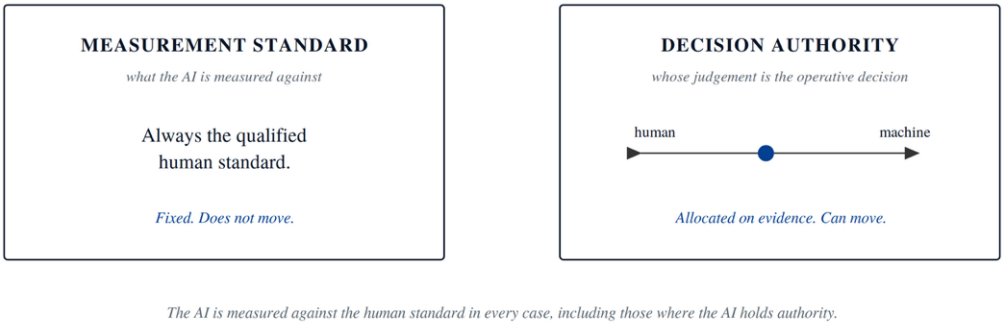

Figure 6. The keystone distinction. The measurement standard is fixed and always human; the decision authority is allocated on evidence and can move between human and machine.

This distinction is also the architecture's central safety property. A layer that allocated authority to machines on any basis other than continuously verified measurement against the human standard would be a mechanism for removing human judgement from consequential decisions. This layer is the opposite: it permits machine-held authority only where a reproducible human standard exists, only where the machine is continuously demonstrated to meet it, and only on classes where meeting it is a sufficient test of fitness.

The human standard here is not the judgement of any single individual, whose fallibility would be a fair objection, but the composed cohort standard of Section 4, whose own reliability is measured rather than assumed. The architecture does not claim that this standard is infallible; Section 8 states plainly where it can be wrong and what the architecture does about it. The claim is narrower and defensible: that a measured consensus of qualified experts, continuously verified and externally anchored, is the best available standard against which to test a machine's fitness to decide, and that no machine may hold authority except by continuously meeting it.

One further matter must be stated plainly. The allocation of decision authority within the architecture does not reallocate legal responsibility. The deploying organisation remains accountable for every decision its systems produce, in every tier, exactly as it is accountable for the decisions of its human staff. What the architecture contributes is not a shield but a record: concurrent, statistical evidence of the standard the system was held to, the basis on which authority was allocated, and the speed with which it was withdrawn when conditions lapsed. An organisation that wanted authority without accountability would find no comfort here.

7.2 The decision-class taxonomy, measured rather than asserted

Authority is allocated by classifying each decision class on a single axis: the degree to which the correct judgement is anchored in verifiable, reproducible evidence, as against the degree to which it is constituted by an irreducibly experiential act of human judgement. At one end, qualified experts converge because the case contains evidence that determines the answer, and disagreement is error; at the other, qualified experts legitimately diverge because the judgement turns on a lived, contextual, or moral act no quantity of evidence resolves, and disagreement is inherent variance, not error.

The decisive property is that the platform measures this axis rather than asserting it. The position of a class is read from the inter-rater reliability the architecture already computes among qualified experts. A class on which independent qualified experts reliably agree admits a reproducible standard; a class on which they persistently and legitimately diverge does not. A reproducible standard, so identified, is one whose answer qualified experts consistently reproduce, which is a claim about agreement and not, on its own, a claim about correctness against the world.

Agreement and observable outcome are distinct axes. Where realised outcomes are observable, they take precedence over inter-rater agreement in this classification, because an outcome can reveal whether persistent disagreement is genuine inherent variance or shared error, and equally whether high agreement is genuine reproducibility or shared bias. The limits of agreement as a proxy for correctness, and the role of outcome anchoring, are set out in Section 8. This classification is computed from the agreement of humans with one another, before any question of the AI's concordance arises, and is therefore independent of how the AI performs. An AI that concords almost perfectly with humans on a class where humans do not agree with each other cannot be granted authority on the basis of that concordance, because the class will have been classified as irreducibly experiential by the human inter-rater measurement that precedes and is independent of the AI's score. This is how the architecture refuses the interpretation it must refuse: the gate on authority is built from human-to-human agreement, which no degree of AI-to-human agreement can open.

The inter-rater signal must, however, be read carefully, because a naive reading is the taxonomy's most likely failure. Low agreement among qualified experts may indicate an irreducibly experiential class, but it may instead indicate a poorly specified case definition, an ambiguous labelling schema, an inadequately briefed cohort, or a sample too small for stability. A low signal is therefore treated as provisional and triggers examination rather than immediate classification.

The competing explanations are excluded by specific, ordered checks: the case definition and labelling schema are reviewed for ambiguity; cohort qualification and briefing are confirmed against the credentialing record; and the sample is grown until the reliability estimate stabilises within its confidence interval. Only disagreement

that persists among demonstrably qualified, adequately briefed experts on an unambiguous definition at a stable sample size is treated as the legitimate inherent variance that marks an experiential class. The classification is thus defensible against its own most likely error, mistaking a badly specified decision for an inherently irreducible one, because the alternative explanations must be ruled out before the experiential reading is accepted.

7.3 The three-tier allocation

The classification yields a three-tier allocation, each tier carrying a different relationship between the AI and the decision, a different evidential burden, and a different continuous condition for the authority to persist.

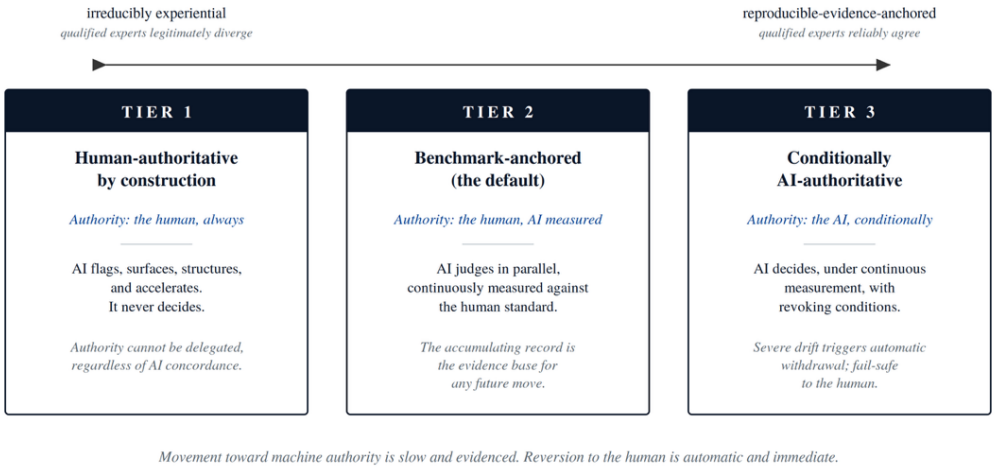

Figure 7. The three-tier allocation of authority on the reproducibility axis. A class falls into a tier by its measured position; movement toward machine authority is slow and evidenced, while reversion to the human is automatic and immediate.

Tier 1, human-authoritative by construction. A class measured as irreducibly experiential is human-authoritative by construction: authority rests with the human not because the AI failed a test, but because the decision is not the kind of thing a machine can hold authority over, regardless of how it performs. The AI's role is to flag, surface, structure, and accelerate, never to decide, and this is enforced as a structural property of the classification so that no configuration error or scope expansion can move a Tier 1 decision into machine hands.

Tier 2, benchmark-anchored. This is the default. The human holds authority and the AI operates in parallel, continuously measured against the human standard. The accumulating concordance record is the evidence base from which a class might, in time and on sufficient evidence, be justified in moving to Tier 3, or be confirmed as properly remaining in Tier 2 indefinitely. Most classes in most deployments will sit in Tier 2, which is the correct and conservative default; an organisation running its entire deployment in Tier 2 is running exactly the verification architecture this paper

describes, with the added assurance that the authority layer has confirmed Tier 2 is where those classes belong.

Tier 3, conditionally AI-authoritative. This is the only tier in which a machine holds authority over a consequential decision, and the conditions are demanding and must hold together: the class is measured as reproducible-evidence-anchored with high and stable inter-rater reliability; the AI's concordance is high, stable, and supported by sufficient accumulated evidence with honest confidence bounds; no drift is present on either surface; and a set of revoking conditions is defined in advance.

The evidential burden is severity-stratified, because an aggregate coefficient can be excellent while saying little about the errors that matter most: disagreements are stratified by the consequence of the case, the rate of severe-consequence disagreement carries its own upper confidence bound, and a class is eligible for Tier 3 only if that upper bound, not the point estimate, sits below the tolerance defined for the class. Where severe-consequence cases are so rare that the bound cannot be driven below tolerance on the accumulated evidence, the class does not graduate, however high its aggregate concordance; insufficient evidence of tail safety is treated as insufficient evidence, not as safety.

Two further conditions complete the tier, and they bind the authority decision to the outcome axis of Section 8. Where realised outcomes are recoverable for the class, outcome evidence is not optional: a class does not graduate while its disagreements with the human standard break against realised outcomes, and the signed concordance of Section 8.3 enters as a required confirming test rather than as a supplement. Where realised outcomes are not recoverable, Tier 3 certifies equivalence to the best externally anchored human standard available and no more, and its justification is explicitly counterfactual. The machine is held to the standard the qualified human cohort sets, measured against it continuously, and withdrawn automatically on any lapse, which is a stronger discipline than the unmeasured human judgement it stands in for, not a weaker one.

The architecture does not claim that the machine decides correctly in these classes, because correctness is unobservable there for human and machine alike; it claims that the machine is held to, and continuously demonstrated to meet, the same standard society already accepts from its qualified professionals, under measurement those professionals are not themselves subject to. Authority is never held outright. It is a continuously renewed grant, affirmed by the live measurement, and any breach of a revoking condition, including severe drift as set out in Section 6, returns authority to the human automatically and immediately.

Losing authority to the human is instant and automatic. Gaining authority for the machine is slow and evidenced. This is the only safe direction for the asymmetry to run, and the architecture is built so that the conservative movement requires no

friction, and the permissive movement carries the full evidential burden of the destination tier.

7.4 Independence and anchor strength, proportionate to delegated authority

The external anchoring of Section 4 and the authority allocation here are connected by a single principle: the strength of the benchmark, in both its independence and the anchor it rests on, scales with the authority being delegated. Where the human retains decision authority, in Tier 2, an internally and externally composed standard with measured inter-rater reliability is sufficient, because the human still owns every decision and the stakes of any residual circularity are bounded. Where the machine is granted authority, in Tier 3, this is not optional: the benchmark must satisfy the external composition floor of Section 4.3 and be demonstrably calibrated to the established external standard, and, in classes where realised outcomes are recoverable, it must additionally draw on the outcome anchoring of Section 8.3, because that is the point at which residual circularity or undetected shared bias would become dangerous.

The progression from an internally composed benchmark, to a fully externally anchored one, to one confirmed against realised outcomes is therefore not a matter of cheaper versus stronger compliance; it is independence and evidential strength scaling with consequence, the same logic a regulator applies when reserving its lightest oversight for the lowest-risk activities.

8 The limits of human equivalence

A governance architecture that measures equivalence to current human judgement must be honest about what that target is and is not. In some contexts equivalence to current human professional judgement is not the right target, because human professionals are sometimes systematically wrong. The cognitive heuristics and biases that distort judgement under uncertainty are long established,[54] they are documented specifically in clinical diagnosis, where anchoring and related biases are associated with diagnostic error,[55,56] and historically embedded bias is well evidenced, for example in the racial disparities that persist in clinical decision-making.[57] An architecture that holds AI to the standard of current human judgement may, in those contexts, be measuring against a ceiling rather than a floor. This section states that limit plainly and sets out what the architecture does about it. The limit is not a marginal caveat; it is a boundary of the entire approach, and it is stated as one.

8.1 Equivalence is not optimality

The concordance coefficient measures equivalence to the qualified human standard. It does not measure optimality. Where the human standard is itself the best available representation of correct practice, equivalence is the right target and the measurement is sufficient. Where the human standard carries systematic bias, equivalence to it certifies that the AI judges as humans judge, including their errors,

and not that the judgement is correct against the world. A high coefficient is evidence that the AI matches the human standard; it is not, on its own, evidence that the human standard is right, and the distinction is stated wherever the coefficient is reported.

This has a consequence that must be confronted directly rather than left implicit. For the large and important class of decisions where realised outcomes are never cleanly recoverable, where the world does not return a verdict, returns it too late to govern anything, or returns a verdict that is itself the matter in dispute, the architecture can certify conformity to the qualified human standard but cannot certify correctness against the world, because no observable ground truth exists to certify against. This is not a deficiency of the architecture; it is a property of the decision. No system, human or machine, can certify correctness where correctness is unobservable. What the architecture does, and what it claims, is to hold the AI to the best externally anchored human standard available and to say openly that this is conformity to that standard, not proven optimality. Stating this boundary is itself a discipline the approach demands.

8.2 Three mechanisms, three reaches

The architecture addresses the limits of human equivalence through three mechanisms of increasing reach. Composition of the cohort with independent, externally endorsed experts catches organisation-level and institutional bias, because the independent experts do not share the deploying organisation's particular blind spots, and divergence between the internal and external cohorts makes that bias visible. External anchoring to certification reference cases catches drift of the cohort away from the codified professional standard, because the reference cases carry an externally established correct answer. Neither catches profession-wide bias: if an entire profession shares a systematic error, both internal and independent experts, drawn from that profession, share it, and the certification standard, set by that profession, may encode it. The only one of these mechanisms that reaches profession-wide bias is anchoring to realised outcomes, developed below, and it is available only where outcomes are recoverable.

The architecture's claim is therefore layered and exact: it cures self-assessment circularity and institutional bias by composition and external anchoring; it surfaces profession-wide bias only where realised outcomes can be observed; and where outcomes cannot be observed it measures conformity to the best externally anchored human standard available and states openly that this is conformity, not proven optimality.

8.3 Outcome anchoring, where the world returns a verdict

Where realised outcomes are recoverable, they are the most powerful evidence available, and the architecture is designed to use them. Where outcomes are recoverable, the concordance coefficient can be signed: each disagreement between the AI and the human standard is resolved against the realised outcome, and the

coefficient records not only the magnitude of agreement but the direction in which disagreements break relative to truth. This is the mechanism that can catch the case where the AI concords well with humans but both, or the AI, diverge from what actually happened. It enters the authority allocation as confirming evidence, strengthening or weakening a placement established on the human standard, and it is bounded so that it sharpens but never solely carries an allocation, because it is absent on the classes where outcomes cannot be observed.

RETROSPECTIVE ANCHORING AND COLD-START AUTHORISATION

The latency that makes live outcomes slow does not apply to the past. An organisation deploying a new agent typically holds an archive of historical cases whose outcomes are already settled. Running those cases through the new agent now produces, for each case, three values: the AI's judgement, the recorded human decision, and the realised outcome. This triangulation establishes, before the agent makes a single live decision, whether its disagreements with the historical human decisions tended to be vindicated by reality, providing pre-deployment authorisation evidence with no production exposure.

The limits are stated honestly. History anchors the future only while the conditions that generated it still hold, so the retrospective pass sets an authorisation prior and never grants standing authority. Archived human decisions are typically a single-reviewer record, not the inter-rater-verified benchmark, so the pass informs but does not replace the classification. And where a past decision foreclosed its alternative, the outcome is censored: a declined application never reveals whether it would have succeeded, so the architecture identifies which classes admit clean retrospective outcomes and which are confounded and reports the weaker evidence honestly rather than a confident result the data cannot support.

9 Reach, economics, and deployment

9.1 The reach of the architecture: rendering judgement assessable

A reasonable reader may suspect that the architecture applies only to decisions clean enough for a direct statistical comparison, and that the messier judgements making up much of consequential professional work lie outside its reach. The obstacle is not where that suspicion places it. A messy, unstructured human input and a complex reasoning process are not obstacles: the concordance engine requires only that a qualified human assessing the same case can be compared to the AI on the judgement each reaches. Where the AI’s output resolves to a position, a determination, a rating, or a structured assessment, the existing machinery governs it. The genuine obstacle is narrower: the decision whose output is itself open-ended, where the AI returns judgement-laden prose that two qualified professionals would word differently while both being correct, and where no defined output space exists over which agreement can be computed.

The resolution is the move the regulated professions have always made. A clinician's holistic impression is unstructured and experiential, yet the profession governs it by rendering it into structured instruments and rated scales so that the judgement can be examined and held to a standard. Above-the-loop governance inherits this discipline. An open-ended output is governed not by comparing the prose but by extracting the assessable judgements the prose embodies, the determinations it makes and the dimensions it expresses, and measuring concordance on that rendered representation against the same representation extracted from a blinded qualified human's assessment of the same case. The wording is never the unit of comparison; the judgement the wording embodies is.

> *The framework does not require the decision to be simple. It requires the judgement to be rendered into an assessable representation, as the regulated professions render holistic judgement into structured instruments in order to govern it. What can be rendered assessable can be governed; what cannot be rendered assessable cannot be governed by anyone and stays human.*

The taxonomy of Section 7 then operates within a single complex output, dimension by dimension. When a judgement is decomposed, its dimensions do not all sit at the same point on the axis. Whether a presenting concern was captured, or a required follow-up determined, are dimensions on which qualified professionals reliably agree, reproducible and governable toward authority; whether a response carried the right empathy, or the irreducibly human read of an individual was sound, are dimensions on which professionals legitimately diverge, held in Tier 1 where the AI assists and the human decides. A complex output is therefore governed not by a single allocation but dimension by dimension, the reproducible dimensions eligible for machine authority and the experiential dimensions reserved to the human; consistent with Section 4.2, any aggregation of per-dimension agreement exists for reporting only and never carries a drift determination or an authority allocation.

This reach is not free, and the cost is a validity burden that must be stated. The rendering of an open-ended judgement into an assessable representation is itself a load-bearing component: a rendering that extracted the wrong dimensions would measure concordance precisely and confidently on the wrong things. The decomposition must therefore be validated against qualified human agreement, ideally confirmed by the endorsing professional body, that those are the right dimensions, that they capture the consequential content, and that they omit nothing material. Until a rendering for a class of complex output has met that burden, the architecture cannot claim to govern that class and says so. The capability is real, it is bounded, and the boundary is stated.

9.2 The economics of governance at scale

Above-the-loop governance is not only the more rigorous approach; under the stability conditions of Section 5.1 it is the approach whose economics remain viable

as AI scales. In-the-loop governance implemented at the rigour required to support defensible population-level claims requires human review that grows linearly with AI volume, and the staffing this implies exceeds the available pool of qualified reviewers in many domains. Above-the-loop governance requires human judgement only on a statistically representative, stratified, risk-weighted sample, so the cost of human judgement scales with statistical sufficiency and exposure tolerance rather than with AI volume, and the cost per decision governed falls as volume grows under stable conditions. The contrast is structural: per-decision review is linear in volume by construction, while sampled verification is sub-linear by construction, an economic property it shares with every sampling-based assurance regime and which per-decision review can never acquire. The economic constraint is as binding as the regulatory one, and this architecture addresses both. Senior decision-makers are therefore not choosing between safer and faster approaches; they are choosing between approaches that can scale with their AI ambitions and approaches that structurally cannot.

9.3 Deployable now, with regulatory acceptance distinguished from technical readiness

The architecture uses infrastructure that already exists. Medical registration boards, bar associations, accountancy and audit bodies, and financial-services credentialing regimes have, over decades and in some cases centuries, established what qualified professional judgement looks like, how it is verified, and how it is remediated. The architecture inherits this infrastructure rather than inventing a new institution: the human cohort is composed of professionals whose standing those institutions already govern, and the independent experts and certification reference cases are drawn from the same bodies. The concordance engine sits between this established human infrastructure and the AI, applying methods established in clinical research and quality science.

A distinction must be drawn that looser framings leave implicit. The technical ability to deploy the architecture is genuinely fast, because nothing new needs to be invented at the institutional level. Regulatory acceptance of the architecture as a substitute for, or complement to, existing human-oversight obligations is a separate matter and will proceed at the pace of each jurisdiction's supervisory process. The architecture is deployable, in the technical sense, in months rather than years; the regulatory recognition of what that deployment satisfies will develop over a longer horizon, and organisations should plan for both timelines rather than conflating them. Nothing in this paper should be read as a claim that satisfying its criteria discharges any legal obligation; that determination belongs to legislators, regulators, supervisors, and courts.

Recent supervisory signals nonetheless indicate the direction, and they are quoted precisely because their wording matters. In April 2026 the Australian Prudential Regulation Authority, following a targeted review of large banks, insurers, and superannuation trustees, found that assurance practices are not keeping pace with the scale, speed and complexity of AI; observed "reliance on point in time and sample

based assurance methods, despite these methods being ill suited to probabilistic models that learn, adapt and degrade over time"; reported that "few entities had continuous validation or monitoring in place to detect issues such as model drift, bias, failure modes, or control breakdowns in a timely manner"; stated that monitoring should be continuous and proportionate to the criticality of the use case; and held stronger supervisory action and, where appropriate, enforcement in reserve for entities that fail to manage AI risk.[52]

One clarification is owed, because the supervisory criticism of "sample based" methods could be carelessly read against any architecture that samples. What APRA criticised is static, point-in-time sampling: an assurance snapshot taken once and left to age against a system that continues to change. What this architecture provides is the opposite of that and the thing the review found missing: continuous statistical sampling under an explicit drift-detection design, in which the sample is perpetually refreshed, the detection latency is a governed design quantity, and model drift is the named object of the measurement. The post-market monitoring instruments surveyed in Section 3.4, from predetermined change control plans to Article 72 of the EU AI Act, point the same way: the obligation to monitor continuously is arriving across jurisdictions, and what remains unspecified is the substrate by which it is performed.

9.4 Security, privacy, and verifiable credentials

The architecture operates inside the deploying organisation. The internal experts are the organisation's own personnel, operating under the employment, confidentiality, and professional-conduct frameworks that govern any other consequential work they perform, and the case material that flows through the internal loop does not leave the organisation's regulatory and contractual perimeter. This matters for frameworks such as the Australian Prudential Regulation Authority's CPG 235 on data risk and CPS 230 on operational risk,[53] and for health-sector privacy obligations. Where independent external experts participate, they do so under endorsement by the relevant professional or regulatory body, under the engagement and payment model of Section 4.3, and under defined confidentiality obligations, blinded to the client, so that the independence that gives their judgement its value is preserved without exposing client-identifying material. The independence of the external pool is inherited from the endorsing body rather than conferred by the platform operator: the bodies nominate or accredit the experts and supply the reference cases that keep them calibrated, and the operator runs the matching, the blinded allocation, the fixed-fee administration, and the measurement. That boundary, together with the structural separation of fees from judgements, is what makes the independence claim defensible rather than a more sophisticated form of the circularity it exists to cure.

10 The framework for evaluating above-the-loop systems

The framework introduced in Section 2 is restated here in full, with the guidance for applying each criterion. Each criterion traces to a documented failure mode or regulatory expectation identified in this paper: the contamination and anchoring evidence of Section 1.2 (criterion 1), the self-assessment circularity of Section 4.3

(criterion 2), the estimation and detection requirements of Section 5 (criteria 3 to 5), the response and authority gaps of Sections 6 and 7 (criteria 6 and 7), the equivalence limits of Section 8 (criterion 8), and the operational and supervisory expectations of Section 9 (criteria 9 and 10). It is offered to the field as a contribution to the integrity of the category, and it applies to every system claiming to deliver above-the-loop governance. A system that satisfies all criteria delivers the category; a system that satisfies few does not, regardless of the language it uses to describe itself.

10.1 Architectural criteria

i. **Independence of the systems** Verify that human experts are blinded at the data layer, not merely in the interface, and that the AI does not receive the human benchmark as an input. Independence of the per-case judgement act, not of all shared knowledge, is the property; it can be inspected.

ii. **An externally anchored, representative human standard** Verify that the cohort carries credentials appropriate to the case classes, that it meets statistical sufficiency, that it includes independent experts endorsed by the relevant external bodies and engaged under an influence-resistant payment model, and that it is calibrated against established certification reference cases where these exist. Where they do not, verify that the weaker anchor is reported honestly.

iii. **Risk-weighted, information-valued sampling with design-weighted estimation** Verify that sampling incorporates consequence weighting and information value, grounded in published literature rather than proprietary heuristics that cannot be independently evaluated, and that population-level estimates are computed with inverse-probability design weighting under versioned routing policies.

iv. **Continuous inter-rater reliability** Verify that inter-rater reliability is calculated and reported wherever more than one expert judges a case, using established statistics interpretable against accepted thresholds, and that low-agreement classes are examined for specification and cohort defects before being read as inherently variant.

v. **Auditable equivalence measurement** Verify that the equivalence record is preserved, externally inspectable, documented to a level supporting independent reproduction, that confidence bounds are honest expressions of the evidence, that prevalence is monitored and reported alongside chance-corrected statistics, that monitored values carry anytime-valid uncertainty statements that remain correct under continuous inspection, and that detection thresholds account for multiplicity and distinguish material from merely detectable divergence.

10.2 Operational criteria

vi. **A defined response to detected drift** Verify that detection is matched by a graded, accountable, time-bound response, that conditional machine

authority is withdrawn automatically where drift is severe, and that remediation is verified by measurement rather than attestation.

vii. **Evidence-based allocation of authority** Verify that the system classifies decision classes on a defensible, empirically anchored axis, allocates authority on that classification rather than by assertion, builds the authority gate from human-to-human agreement, applies severity-stratified evidence requirements before any grant of machine authority, and fails toward the human.

viii. **Honest treatment of human-equivalence limits** Verify that the system distinguishes conformity from optimality, states which biases its mechanisms can and cannot reach, and anchors to realised outcomes where these are recoverable.

ix. **Consistent, reliable operation at enterprise scale** Verify that the system produces consistent governance output across deployments without bespoke engineering for each, operates reliably under production conditions, and demonstrably preserves the sub-linear scaling property in operation under stable conditions rather than asserting it in theory.

x. **Deployment without new bureaucracy, with verifiable credentials** Verify that the system deploys using existing professional infrastructure, that experts participate without disruption to their practice, and that current, documented credentials are recorded for every expert, internal and independent, throughout their participation.

A system satisfying most but not all criteria is at risk of failing where it falls short; the importance of each criterion should be weighted according to the regulatory and operational realities of the deployment context. The maturity of the category will depend on the rigour with which the criteria are applied.

11 The architecture for the next decade

Above-the-loop governance is offered as the architectural layer required to verify, continuously and at scale, that AI outputs, primarily judgements of consequence, in high-risk and regulated contexts remain equivalent to the qualified human professional standard, and to govern, on evidence, which decisions a machine may appropriately hold. It is not a replacement for pre-deployment safety work, calibration cycles, or in-loop and on-loop oversight; it is the complementary layer that fills the verification and authority gaps those layers leave between them at modern scale.

The case does not rest on novelty. It rests on the application of statistical methodology that has underwritten consequential decision-making for more than half a century, applied to a class of decision-maker that has not previously existed, and extended in three disciplined ways: by anchoring the human benchmark to externally legitimate standards so that an organisation does not mark its own work; by governing the appointment of authority so that delegation to machines is

evidenced, conditional, and revocable; and by defining the response that follows detection so that an alarm becomes a governed obligation.

The paper has positioned this contribution within the literatures nearest to it: the scalable oversight and AI control programmes of AI safety research, whose deployment-time gap it addresses under an explicitly stated non-adversarial scope, and the post-market monitoring regimes of regulated industries, whose obligations it supplies a substrate for discharging. The paper has also named its lineage: the settlement by which society takes products it will never individually test, on the strength of phased authorisation, sampled release, continuous surveillance, and revocable authority, is the settlement constructed here for deployed AI. Where the world reveals outcomes, the architecture anchors its claims to reality; where it does not, it stands openly on the externally anchored human standard and says so. Throughout, the paper has stated its assumptions and its limits rather than concealing them, on the view that a governance method which blurred the distinction between what it proves and what it asserts would reproduce the failure it exists to prevent.

> *Above-the-loop governance preserves qualified human professional judgement as the standard against which AI is held, while allowing AI to operate at the scale its applications require, and proves, continuously and on evidence, that each decision is held by the party that should hold it.*

The next decade of AI deployment will be governed by the architectures established in the next several years. The choice before enterprises, governments, and regulators is whether that architecture will remain one of the existing positions on its own, with their established limitations, or whether the additional verification and authority layer described here will be adopted as the complementary architecture that closes the gap. What the field cannot continue to accept is governance that exists in language but not in substance; the cost of maintaining the appearance of oversight without its operational reality is paid by the individuals and communities whose lives and livelihoods are affected by the decisions AI is now empowered to make.

Competing interests

The author founded and serves as Director of Trakked Pty Ltd, which operationalises the architecture described in this paper as a deployable governance system. The evaluation framework set out in Sections 2 and 10 is offered as a neutral instrument applicable to every provider claiming the category, and the author holds that any system he builds should be tested against it as rigorously as any other. Readers should weigh the argument of this paper on its merits and against the published framework. A provisional patent application covering certain methods referred to in this paper was filed in May and updated in June 2026 and is pending.

Funding

This work received no external funding.

References

1. McKinsey & Company (2025). The State of AI in 2025: Agents, Innovation, and Transformation. McKinsey Global Survey conducted 25 June to 29 July 2025; 1,993 respondents across 105 nations. Published 5 November 2025.

2. Gartner (2025). Gartner Predicts 40% of Enterprise Applications Will Feature Task-Specific AI Agents by 2026, Up from Less Than 5% in 2025. Press release, 26 August 2025.

3. Maslej, N., Altman, R., Brynjolfsson, E., et al. (2026). Artificial Intelligence Index Report 2026. Stanford Institute for Human-Centered Artificial Intelligence, Stanford University, April 2026.

4. Financial Conduct Authority and Bank of England (2024). Artificial Intelligence in UK Financial Services - 2024. Third joint survey, 118 firms; reporting that approximately 75% of firms use AI, that 55% of AI use cases involve some automated decision-making, and that only 2% are fully autonomous. Published 21 November 2024.

5. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act). Article 14: Human Oversight, including Article 14(4) on automation bias and Article 14(5) requiring separate verification by at least two competent natural persons for the remote biometric identification systems referred to in Annex III point 1(a); Article 26 on the obligations of deployers of high-risk AI systems; and Article 72 on post-market monitoring by providers. Official Journal of the European Union, 12 July 2024.

6. Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association, 19(1), 121-127; and Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: a systematic review. Journal of the American Medical Informatics Association, 24(2), 423-431.

7. Horowitz, M. C., & Kahn, L. (2024). Bending the automation bias curve: a study of human and AI-based decision making in national security contexts. International Studies Quarterly, 68(2), sqae020.

8. Rosbach, E., Ammeling, J., Ganz, J., Bertram, C. A., Conrad, T., Riener, A., & Aubreville, M. (2026). Stuck on suggestions: automation bias, the anchoring effect, and the factors that shape them in computational pathology. Journal of Machine Learning for Biomedical Imaging (MELBA); arXiv:2603.11821, March 2026.

9. Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775-779.

10. Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. Computer Law & Security Review, 45, 105681.

11. Bansal, G., Wu, T., Zhou, J., et al. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.

12. Bank of England (2026). Artificial Intelligence Consortium minutes, 9 February 2026, recording members' views; and Summary of AI roundtables, published 16 February 2026, recording participants' views under the Chatham House Rule. The views recorded in both documents are those of members and participants and are stated not to reflect the views of the Bank of England or the Financial Conduct Authority.

13. Gartner (2025). Gartner Predicts Over 40% of Agentic AI Projects Will Be Cancelled by End of 2027. Press release, 25 June 2025; reporting both that over 40% of agentic AI projects will be cancelled by the end of 2027, citing escalating costs, unclear business value, or inadequate risk controls, and that at least 15% of day-to-day work decisions will be made autonomously through agentic AI by 2028, up from 0% in 2024.

14. Strauss, I., Moure, I., O'Reilly, T., & Rosenblat, S. (2025). Real-World Gaps in AI Governance Research. AI Disclosures Project, Social Science Research Council; arXiv:2505.00174.

15. Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency (2011). Supervisory Guidance on Model Risk Management. SR Letter 11-7, 4 April 2011.

16. Prudential Regulation Authority (2023). Model risk management principles for banks. Supervisory Statement SS1/23, Bank of England, 17 May 2023.

17. Scharre, P., & Horowitz, M. C. (2015). Meaningful Human Control in Weapon Systems: A Primer. Center for a New American Security.

18. Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: a philosophical account. Frontiers in Robotics and AI, 5, 15.

19. Fink, M. (2025). Human Oversight under Article 14 of the EU AI Act. Social Science Research Network, February 2025.

20. Regulation (EU) 2016/679 (General Data Protection Regulation), Article 22; and Court of Justice of the European Union, SCHUFA Holding (Scoring), Case C-634/21, judgment of 7 December 2023.

21. McKinsey & Company (2025). The agentic organization: Contours of the next paradigm for the AI era. September 2025.

22. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv:1606.06565.

23. Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. arXiv:1805.00899.

24. Christiano, P., Shlegeris, B., & Amodei, D. (2018). Supervising strong learners by amplifying weak experts. arXiv:1810.08575.

25. Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). Scalable agent alignment via reward modeling: a research direction. arXiv:1811.07871.

26. Cotra, A. (2021). The case for aligning narrowly superhuman models (proposing the sandwiching paradigm); and Bowman, S. R., et al. (2022). Measuring progress on scalable oversight for large language models. arXiv:2211.03540.

27. Burns, C., et al. (2023). Weak-to-strong generalization: eliciting strong capabilities with weak supervision. arXiv:2312.09390.

28. Engels, J., et al. (2025). Scaling laws for scalable oversight. arXiv:2504.18530.

29. Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2023). AI control: improving safety despite intentional subversion. arXiv:2312.06942.

30. Bhatt, A., Rushing, C., Kaufman, A., et al. (2025). Ctrl-Z: controlling AI agents via resampling. arXiv:2504.10374.

31. Korbak, T., Clymer, J., Hilton, B., Shlegeris, B., & Irving, G. (2025). A sketch of an AI control safety case. arXiv:2501.17315.

32. Shah, R., et al. (2025). An approach to technical AGI safety and security. arXiv:2504.01849; identifying the use of limited human supervision capacity to oversee large AI deployments as an open requirement.

33. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), 1-37.

34. Embi, P. J. (2021). Algorithmovigilance: advancing methods to analyze and monitor artificial intelligence-driven health care for effectiveness and equity. JAMA Network Open, 4(4), e214622.

35. United States Food and Drug Administration, Health Canada, and the Medicines and Healthcare products Regulatory Agency (2023). Predetermined Change Control Plans for Machine Learning-Enabled Medical Devices: Guiding Principles; drawing on Good Machine Learning Practice guiding principle 10, that deployed models are monitored for performance and re-training risks are managed; and United States Food and Drug Administration (2025). Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations. Draft guidance, January 2025.

36. International Organization for Standardization and International Electrotechnical Commission (2023). ISO/IEC 42001:2023, Information

technology - Artificial intelligence - Management system. Published December 2023.

37. International Organization for Standardization and International Electrotechnical Commission (2023). ISO/IEC 23894:2023, Information technology - Artificial intelligence - Guidance on risk management.

38. National Institute of Standards and Technology (2023, with companion updates through 2025). AI Risk Management Framework (AI RMF 1.0), NIST AI 100-1, organising governance into the Govern, Map, Measure, and Manage functions, with continuous monitoring a core requirement of the Measure function; with the associated Generative AI Profile, NIST AI 600-1 (2024).

39. Cochran, W. G. (1977). Sampling Techniques (3rd ed.). John Wiley & Sons. The standard reference for sample-size determination and stratified sampling.

40. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46.

41. Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample-size requirements. Physical Therapy, 85(3), 257-268; McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia Medica, 22(3), 276-282; and Krippendorff, K. (2018). Content Analysis: An Introduction to Its Methodology (4th ed.). SAGE.

42. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174. Alternative banding conventions are proposed in McHugh (2012), reference 40.

43. Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology, 43(6), 543-549; and Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. Journal of Clinical Epidemiology, 46(5), 423-429.

44. Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology, 61(1), 29-48.

45. Wald, A. (1945). Sequential tests of statistical hypotheses. Annals of Mathematical Statistics, 16(2), 117-186; Howard, S. R., Ramdas, A., McAuliffe, J., & Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. The Annals of Statistics, 49(2), 1055-1080; Ramdas, A., Grünwald, P., Vovk, V., & Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. Statistical Science, 38(4), 576-601; and Grünwald, P., de Heide, R., & Koolen, W. M. (2024). Safe testing. Journal of the Royal Statistical Society Series B, 86(5), 1091-1128.

46. Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379-423 and 27(4), 623-656.

47. Settles, B. (2009). Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison.

48. United States Food and Drug Administration (2013). Guidance for Industry: Oversight of Clinical Investigations - A Risk-Based Approach to Monitoring. Center for Drug Evaluation and Research.

49. Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47(260), 663-685.

50. Page, E. S. (1954). Continuous inspection schemes. Biometrika, 41(1/2), 100-115. The foundational paper on cumulative-sum (CUSUM) control charts.

51. Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 327(8476), 307-310.

52. Australian Prudential Regulation Authority (2026). Letter to industry on artificial intelligence, 30 April 2026, signed by Member Therese McCarthy Hockey, reporting findings from a targeted review of large banks, insurers, and superannuation trustees conducted in late 2025; quoted passages from the published letter and its attachment, Observations for Executive Management.

53. Australian Prudential Regulation Authority. Prudential Practice Guide CPG 235: Managing Data Risk; and Prudential Standard CPS 230: Operational Risk Management, which commenced on 1 July 2025, with transitional arrangements for pre-existing service provider contracts ending, and targeted amendments commencing, on 1 July 2026. APRA Prudential Handbook.

54. Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. Science, 185(4157), 1124-1131.

55. Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. Academic Medicine, 78(8), 775-780.

56. Saposnik, G., Redelmeier, D., Ruff, C. C., & Tobler, P. N. (2016). Cognitive biases associated with medical decisions: a systematic review. BMC Medical Informatics and Decision Making, 16, 138.

57. Hoffman, K. M., Trawalter, S., Axt, J. R., & Oliver, M. N. (2016). Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. PNAS, 113(16), 4296-4301.