Article

Who Signs the Opinion? Agentic AI, Accountability, and the Delegation Boundary in Actuarial Work

Ortopah Kojo *

Posted Date: 29 July 2026

doi: 10.20944/preprints202607.2097.v1

Keywords: agentic AI; actuarial function; professional accountability; loss reserving; insurance regulation; AI governance; stat-utory sign-off; human oversight

 doi:10.20944/preprints202607.2097.v1

Article

Who Signs the Opinion? Agentic AI, Accountability, and the Delegation Boundary in Actuarial Work

Ortopah Kojo

Actuary and Independent Researcher; okbotchey@orbotrice.com

Abstract

The insurance industry is moving rapidly from generative artificial intelligence as a drafting assistant to agentic artificial intelligence as an autonomous performer of technical work, including loss reserving, pricing and capital analysis. This shift collides with a foundational institution of insurance regulation: the statutory opinion of a named, professionally accountable actuary. Existing professional standards define the actuary's responsibility for work performed by others in terms of supervision and review, concepts developed for human delegates whose reasoning can be interrogated. This paper asks what those concepts mean when the delegate is an artificial agent whose work is fluent, fast, voluminous and only partially reproducible. Drawing on the regulatory architecture of the actuarial signature across jurisdictions, on the documented failure modes of large language models, and on the human-factors literature on automation complacency, the paper argues that the tacit assumptions underlying professional reliance standards, interrogability, error legibility and normative alignment, fail for artificial delegates. It develops three conditions under which professional sign-off on agent-produced work remains meaningful: reproducibility of the quantitative core, traceability of every material judgement to an identifiable locus, and contestability, meaning the reviewing actuary's practical ability to challenge and override the agent before the opinion is issued. The paper further proposes a five-level delegation hierarchy for actuarial artificial intelligence, a reference architecture that separates deterministic computation from linguistic interpretation, and an evaluation protocol through which the three conditions can be tested against any deployed system. The central claim is that the accountability boundary does not break where the technology fails; it breaks where the technology succeeds so smoothly that review quietly degrades into ratification. Implications are drawn for supervisors, for the professional bodies whose reliance standards require amendment, and for capacity-constrained markets where agentic tools are most attractive and reviewing capacity is thinnest.

Keywords: agentic AI; actuarial function; professional accountability; loss reserving; insurance regulation; AI governance; statutory sign-off; human oversight

1. Introduction

Every regulated insurance market rests on a quiet institutional bargain. Policyholders and supervisors cannot themselves verify whether an insurer's technical provisions are adequate, whether its prices are sustainable or whether its capital position is sound. In place of direct verification, the system relies on a named professional, variously called the appointed actuary, the responsible actuary, the actuarial function holder or the signing actuary, who attaches a personal, sanctionable signature to an opinion. The signature is not a formality. It is the mechanism by which diffuse technical uncertainty is converted into concentrated personal accountability, and it is enforced by professional discipline, statutory liability and, in most jurisdictions, the credible threat of losing the licence to practise.

 doi:10.20944/preprints202607.2097.v1

That bargain was struck in a world where actuarial work, however computerised, was performed by humans. When the appointed actuary relied on the work of others, professional standards told her what reliance required: understand the methods, review the assumptions, test the results, and remain able to defend every material judgement as her own. The delegate was a junior actuary, an analyst or an external consultant, in every case a person whose reasoning could be interrogated in ordinary language, whose errors followed recognisably human patterns, and who was, in the last resort, subject to the same professional norms.

Between 2024 and 2026 the character of the delegate began to change. Large language models moved from drafting emails to performing technical work. Insurers now deploy agentic systems, meaning artificial intelligence that plans, selects tools, executes multi-step analytical workflows and returns finished technical products with limited human involvement in intermediate steps. The 19th annual emerging risks survey conducted jointly by the Society of Actuaries Research Institute and the Casualty Actuarial Society (2026) reports that senior insurance executives now rank adverse artificial intelligence outcomes among the top emerging risks facing their organisations, ahead of many traditional insurance perils. Supervisors have responded with instruments such as the model bulletin on the use of artificial intelligence systems by insurers coordinated by the National Association of Insurance Commissioners (2023), the European Union's Artificial Intelligence Act with its high-risk classification of certain insurance applications (European Parliament and Council, 2024), and Swiss supervisory guidance on governance and risk management when using artificial intelligence (FINMA, 2024).

Yet almost the entire regulatory and academic conversation addresses the institution: governance committees, model inventories, bias testing, documentation standards. Remarkably little addresses the person. When an artificial agent performs the reserving analysis, selects the development factors, chooses between the chain-ladder and Bornhuetter-Ferguson indications and drafts the report, who is responsible for the opinion, and what must that person actually have done for the responsibility to be real rather than ceremonial? The question is not hypothetical. The economics of agentic tooling are most compelling precisely in the settings where independent human review capacity is thinnest: small and mid-sized insurers in mature markets, and entire insurance sectors in emerging markets where qualified actuaries number in the dozens rather than the thousands.

This paper addresses the question conceptually, and states its method and its grounding plainly. The argument proceeds by analysis of three bodies of material: the statutory and professional architecture of the actuarial signature across selected jurisdictions; the documented failure modes of large-language-model systems, in particular fabricated content and unfaithful self-explanation; and the human-factors literature on automation complacency, which describes how human oversight decays over reliable automation. The analysis is additionally informed by the author's own practice designing, deploying and operating production actuarial artificial intelligence tools for reserving and pricing, a position disclosed fully in the declarations; that practice motivates the framework and supplies illustrative failure archetypes, but the paper claims no systematic empirical findings, and Section 7 instead specifies the evaluation protocol through which the framework's conditions can be tested empirically, by the author or by others, as future work. The empirical scope of the illustrations is general insurance reserving; life valuation, capital modelling and pricing lie outside the paper's illustrative reach, a boundary revisited in Section 10.

The paper makes four contributions. First, it reframes the actuarial artificial intelligence debate from institutional governance to personal professional accountability, connecting the emerging governance literature with the older and largely separate literature on actuarial professionalism and reliance on the work of others. Second, it introduces a three-condition test, reproducibility, traceability and contestability, that specifies what review must mean when the reviewed party is an artificial agent, and derives each condition from a documented property of the underlying technology or of human oversight behaviour. Third, it develops a five-level delegation hierarchy for actuarial artificial intelligence, analogous in spirit to the levels of driving automation, which gives regulators and appointed actuaries a shared vocabulary for describing how much of the actuarial control cycle has actually been delegated. Fourth, it distils these into a reference architecture and a sign-off readiness

 doi:10.20944/preprints202607.2097.v1

assessment that an actuarial function, an audit committee or a supervisor can apply to any agentic actuarial system.

The argument proceeds as follows. Section 2 characterises the agentic turn in actuarial work and distinguishes it from earlier waves of actuarial automation. Section 3 reviews the regulatory and professional architecture of the actuarial signature and shows that its reliance concepts presuppose a human delegate. Section 4 develops the conceptual framework: the accountability boundary, the three conditions and the delegation hierarchy. Section 5 grounds each condition in the documented properties of the technology and of human oversight. Section 6 translates the conditions into a reference architecture for accountable agentic actuarial systems. Section 7 specifies the evaluation protocol. Section 8 presents the sign-off readiness assessment. Section 9 discusses implications for supervisors, professional bodies and capacity-constrained markets. Section 10 acknowledges limitations, and Section 11 concludes.

2. The Agentic Turn in Actuarial Work

2.1. Three Waves of Actuarial Automation

Automation is not new to actuarial work, and the accountability question posed here can only be understood against what came before. It is useful to distinguish three waves. The first wave, from roughly the 1980s to the 2000s, was computational automation: spreadsheets and specialised reserving and pricing software removed manual calculation. The actuary specified every method, every assumption and every judgement; the machine merely executed arithmetic faster. Accountability was untouched because the machine made no judgements. The second wave, from roughly the 2010s, was statistical automation: generalised linear models, gradient boosting and related machine learning techniques began to make implicit judgements, selecting variable weights and interactions that no human had individually specified (Wüthrich & Merz, 2023). This wave produced the first genuine accountability literature in insurance, focused on explainability, proxy discrimination (Prince & Schwarcz, 2020) and model risk management, and it produced the supervisory model governance frameworks now standard in mature markets, of which the United States federal guidance on model risk management is the archetype (Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, 2011). But even in the second wave, the human retained the workflow: the actuary framed the problem, prepared the data, chose the model class, interpreted the output and wrote the report.

The third wave, agentic automation, is qualitatively different because the machine takes over the workflow itself (Zhu, 2026). An agentic system receives an objective, for example review the adequacy of the booked reserves for this motor portfolio, decomposes it into steps, selects and executes analytical tools, evaluates intermediate results, iterates and returns a finished product including narrative interpretation. The judgements are no longer only implicit weights inside a fitted model; they are explicit, sequential, natural-language decisions about method selection, assumption setting, outlier treatment and the framing of conclusions. In other words, the third wave automates precisely the layer of actuarial work that professional standards assumed would always remain human: the exercise of professional judgement within the actuarial control cycle (Bellis et al., 2010).

2.2. Why Reserving is the Critical test Case

Loss reserving is the natural site for examining the accountability question, for three reasons. First, reserving opinions are the paradigmatic object of statutory actuarial sign-off across jurisdictions: technical provisions under Solvency II and the Swiss Solvency Test, statements of actuarial opinion in the United States, and the actuarial valuations required by insurance acts across African markets. Second, reserving combines mechanisable computation with irreducible judgement in a well-understood way. The chain-ladder method as formalised by Mack (1993) and the method of Bornhuetter and Ferguson (1972) are deterministic given their inputs, and their stochastic extensions are well charted (England & Verrall, 2002), yet the selection of development factors, tail assumptions, a priori loss ratios and the weighting between method indications is classical actuarial judgement. An agentic reserving system therefore automates both the part that is easy to verify and the part that

 doi:10.20944/preprints202607.2097.v1

is not, in a single workflow. Third, reserving errors are consequential and slow to surface. A mis-priced policy reveals itself within a year; an under-reserved portfolio can compound silently for a decade. The accountability structure exists precisely because feedback is too slow for market discipline to substitute for professional discipline.

2.3. The Seduction of Fluency

A recurring theme in this paper is that the accountability risk of agentic systems is not primarily that they fail visibly. Visible failure triggers review. The risk is that they succeed fluently. A modern language-model-based agent produces reserving narratives that are grammatical, confident, correctly formatted, sprinkled with appropriate caveats and structurally indistinguishable from competent human work. Fluency is evidence of competence in a human delegate, because for humans the ability to articulate a judgement is correlated with the ability to make it. For an artificial agent that correlation is broken: fluency is a property of the language interface, not of the underlying analysis.

The reviewing actuary's most practised heuristic, does this read like the work of someone who knows what they are doing, is therefore systematically miscalibrated for agentic work. This inversion, from failure-detection to success-verification, is the core of what changes for the signing actuary, and it motivates the three conditions developed in Section 4.

2.4. Positioning Within the Emerging Literature

A rapidly growing practitioner and thought-leadership literature now addresses agentic artificial intelligence in actuarial work. Professional-body publications, including the Institute and Faculty of Actuaries' work on the agentic model office and the future shape of insurance organisations (Institute and Faculty of Actuaries, n.d.) and the Society of Actuaries' emerging-topics coverage, examine what agents mean for workflow design and organisational structure. The American Academy of Actuaries (2026) has addressed algorithmic accountability at the level of professional ethical standards, framing artificial intelligence use as subject to existing modelling and communication standards, and the International Actuarial Association's artificial intelligence governance framework directs practitioners back to the reliance provisions of ISAP 1 when using vendor and third-party models (International Actuarial Association, 2025). In the adjacent risk literature, agentic capability has been characterised as a continuum of autonomy and delegated authority whose degree is the first-order variable for risk assessment (Zhu, 2026), and practitioner governance commentary has begun to frame the core problem as one of delegation rather than detection (The Actuary, 2026). Vendor and consulting analyses concentrate on efficiency gains and use cases, and the regulatory instruments surveyed in Section 3 address institutional governance. What this literature confirms is that reliance and delegation are recognised as the operative concepts; what it leaves unspecified is their content. To the author's knowledge, as of mid-2026 no published work combines a delegation-level taxonomy for actuarial artificial intelligence with an operationalised test of what personal sign-off over agent-produced work requires. That combination, personal rather than institutional accountability, and conditions derived from the documented properties of the technology rather than from governance principles, is the gap this paper occupies.

3. The Regulatory and Professional Architecture of the Actuarial Signature

3.1. The Signature as an Institution

Across otherwise diverse regulatory regimes, the actuarial signature has a common institutional logic with three elements: a named natural person, a defined statutory object of opinion, and a personal sanction regime. Under the Solvency II framework in the European Union, Article 48 of Directive 2009/138/EC requires insurers to maintain an effective actuarial function that coordinates the calculation of technical provisions, assesses the sufficiency and quality of the underlying data and expresses an opinion on the overall underwriting policy and reinsurance arrangements (European Parliament and Council, 2009). In Switzerland, the Insurance Supervision Act institutionalises the responsible actuary, a named individual who must ensure that technical provisions are adequate and

 doi:10.20944/preprints202607.2097.v1

who bears personal responsibility toward the supervisor (Swiss Confederation, 2004); FINMA's Guidance 08/2024 has since articulated supervisory observations and assessments on governance, inventory and risk classification, data quality, tests and ongoing monitoring, documentation, explainability and independent review where supervised institutions use artificial intelligence (FINMA, 2024). In the United States, the appointed actuary signs an annual statement of actuarial opinion subject to the Actuarial Standards of Practice and the qualification standards of the American Academy of Actuaries. In Ghana, the Insurance Act 2021 (Act 1061) requires licensed insurers to appoint an actuary approved by the National Insurance Commission and to submit actuarial valuations (Republic of Ghana, 2021), a structure broadly mirrored across anglophone African markets under their respective insurance acts.

The common thread is that in every regime the law names a person, not a process. Governance frameworks, model risk policies and validation units support the person, but they do not absorb the person's responsibility. This design is deliberate. Diffuse responsibility invites what the governance literature calls the problem of many hands (Thompson, 1980); the signature is the institutional device that defeats it.

3.2. Reliance on the Work of Others: The Load-Bearing Concept

No signing actuary performs all the work personally, and professional standards have always accommodated this through reliance provisions. International Standard of Actuarial Practice 1 of the International Actuarial Association permits reliance on the work of others where the actuary considers it reasonable to do so, subject to disclosure and to the actuary taking responsibility for the work unless responsibility is explicitly disclaimed and the disclaimer is permitted (International Actuarial Association, 2018). The Financial Reporting Council's Technical Actuarial Standard 100 in the United Kingdom requires that judgements be exercised, communicated and documented so that a technically competent reader can understand them (Financial Reporting Council, 2023). The Actuarial Standards of Practice in the United States, notably those addressing data quality, reliance, actuarial communications and, since 2019, modeling (Actuarial Standards Board, 2019), elaborate what the relying actuary must review and disclose, and the professional bodies of African markets incorporate equivalent conduct standards, typically modelled on International Actuarial Association templates. Notably, when the International Actuarial Association's task force turned to artificial intelligence governance, it directed practitioners back to precisely these reliance provisions for the use of vendor and third-party models (International Actuarial Association, 2025), confirming that reliance is the profession's operative concept for artificial delegates while leaving open the question this paper addresses: what reliance can legitimately consist of when the delegate's properties differ from those the provisions assume.

Read closely, these reliance provisions embed three tacit assumptions about the delegate. First, interrogability: the delegate can explain, in dialogue, why a judgement was made, and the explanation is causally connected to the judgement actually made. Second, error legibility: the delegate's mistakes follow patterns the reviewer has learned to detect, arithmetic slips, misapplied methods, optimistic tail selections, because reviewer and delegate share a common training. Third, normative alignment: the delegate is subject to the same professional duties, so the reviewer can presume good faith and calibrated confidence, and deviations are individually culpable. All three assumptions fail, in whole or in part, for an artificial agent. An agent's natural-language explanation of its own output is a generated artifact that may or may not correspond to the computational path actually taken. Its errors are novel in kind: confident fabrication, silent omission, plausible interpolation, categories with no stable human analogue. And it bears no professional duties at all; every normative property it displays is a design outcome, not an ethical commitment. Section 5 grounds each of these failure claims in the documented literature.

3.3. The Gap in the Emerging AI Governance Instruments

The new supervisory instruments address the institution admirably and the person hardly at all. The model bulletin coordinated by the National Association of Insurance Commissioners (2023) re-

 doi:10.20944/preprints202607.2097.v1

quires insurers to maintain a written artificial intelligence systems program with governance, risk management and internal controls. The European Union's Artificial Intelligence Act imposes provider and deployer obligations, human oversight requirements and documentation duties on high-risk systems, with certain insurance applications explicitly in scope (European Parliament and Council, 2024). Swiss guidance emphasises inventory, risk classification, documentation and independent review at supervised institutions (FINMA, 2024). At the international standard-setting level, the application paper of the International Association of Insurance Supervisors (2025) elaborates governance, risk management and human oversight expectations for insurers' use of artificial intelligence, and the Actuarial Association of Europe (2025) has mapped the Artificial Intelligence Act's requirements onto actuarial ethical and governance responsibilities. These are institutional instruments: they regulate programs, inventories and committees. None of them answers the appointed actuary's question, which is individual and immediate: what must I personally have done, before signing an opinion resting on agent-produced work, for my signature to mean what the statute intends it to mean? The professional bodies, for their part, have issued thoughtful discussion papers on artificial intelligence and actuarial work, but as of mid-2026, and to the author's knowledge, no major actuarial standard-setter has amended its reliance standards to address artificial delegates specifically. This paper aims to supply the missing analytical layer.

4. Conceptual Framework: The Accountability Boundary

4.1. Defining the Boundary

Define the accountability boundary as the frontier within an actuarial workflow up to which the signing actuary can truthfully assert: I have either performed this work or reviewed it in a manner sufficient for me to adopt its judgements as my own. Work inside the boundary is covered by the signature in substance; work outside it is covered only in form. The central claim of this paper is that agentic systems do not move the boundary so much as blur it, because they produce artifacts whose form signals reviewed-ness while their substance resists the review practices the profession actually uses. The task is therefore to specify conditions under which the boundary can be redrawn sharply around agent-produced work.

4.2. Three Conditions for Meaningful Sign-off

The framework proposes that professional sign-off on agent-produced actuarial work is meaningful if and only if three conditions hold. They are individually necessary; Section 5 derives each from a documented property of the technology or of human oversight behaviour, and Section 7 specifies how each can be tested against a deployed system.

Condition one: reproducibility of the quantitative core. Every number that flows into the opinion, reserve indications, development factors, adequacy test statistics, must be reproducible by a deterministic path from identified data inputs, such that the reviewing actuary or an independent party can regenerate it exactly. In architectural terms this requires that the agent's quantitative work be executed by conventional, versioned computational code that the agent invokes, not by the language model's own token generation. A reserve estimate produced inside a language model's forward pass is not reproducible in the required sense even if it happens to be correct, because there is no path from inputs to output that a reviewer can independently traverse. Reproducibility converts the largest part of the review problem back into the familiar second-wave problem of validating deterministic code, which the profession knows how to do.

Condition two: traceability of judgement. Every material judgement embedded in the output, the selection among method indications, the treatment of an anomalous diagonal, the characterisation of adequacy, must be traceable to an identifiable locus: a prompt instruction, a configuration parameter, a data feature or an explicit agent decision recorded in an execution log. Traceability is weaker than explainability and deliberately so. The framework does not require that the agent's self-explanations be faithful accounts of its internal computation, a requirement current systems cannot meet, as Section 5.2 shows. It requires only that the reviewing actuary be able to locate where each judge-

 doi:10.20944/preprints202607.2097.v1

ment entered the workflow, so that she can evaluate the judgement on its merits rather than evaluating the agent's rhetoric about it.

Condition three: contestability before issuance. The reviewing actuary must have the practical ability, in time, tooling and expertise, to challenge any judgement, substitute her own, and have the substitution propagate correctly through the final product before the opinion is issued. Contestability is the condition most likely to be silently violated in practice, not by system design but by workload economics: an agent that produces in minutes what previously took weeks creates organisational pressure to review at the speed of production. Where the effective review window shrinks below the time required to exercise conditions one and two, contestability has failed even if the interface offers an edit button. Contestability is therefore a property of the sociotechnical deployment, not of the software alone.

4.3. A Five-Level Delegation Hierarchy for Actuarial Artificial Intelligence

Debates about artificial intelligence and the actuarial function are frequently confused because participants imagine different degrees of delegation. The premise that degree of delegated authority, rather than the presence of artificial intelligence as such, is the first-order variable is shared by the emerging risk literature, which characterises agentic capability as a continuum of autonomy and delegated authority for underwriting and risk-assessment purposes (Zhu, 2026). Borrowing the rhetorical structure, though not the content, of the levels of driving automation, the framework distinguishes five levels tailored to the actuarial control cycle. Level 0, computational assistance: the system executes calculations specified entirely by the actuary; all judgement is human. This is the first wave and poses no novel accountability issue. Level 1, drafting assistance: the system produces text, summaries or code under continuous human direction; judgements are human, articulation is machine-assisted. Level 2, supervised analysis: the system executes a complete analytical workflow, method application through draft interpretation, but every judgement point is surfaced for explicit human decision before the workflow proceeds. Level 3, reviewed autonomy: the system executes the complete workflow including judgements, and the human reviews the finished product with the ability to contest and override before issuance. Level 4, delegated authority: the system's output enters downstream use, booking, filing, disclosure, without item-level human review, subject only to periodic or exception-based oversight.

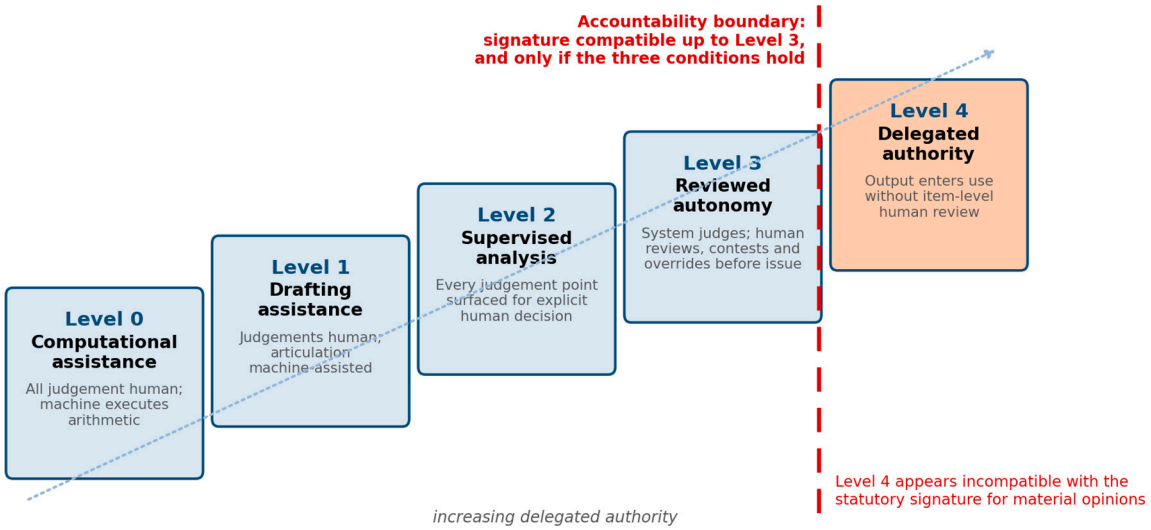

Figure 1. The five-level delegation hierarchy for actuarial artificial intelligence and the position of the accountability boundary.

Figure 1 depicts the hierarchy and the position of the accountability boundary. The hierarchy does two kinds of work. Descriptively, it gives supervisors and boards a vocabulary for asking the only question that matters, namely at what level is this system actually operating, as opposed to the

 doi:10.20944/preprints202607.2097.v1

level claimed in the governance documentation. The gap between claimed Level 2 and actual Level 3, or claimed Level 3 and actual Level 4, is where accountability quietly evaporates, and Section 5.3 explains the behavioural mechanism of the drift.

Normatively, the framework's position, defended in Section 9, is that under current statutory regimes Level 4 appears incompatible with the institution of the actuarial signature for material opinions: a signature over unreviewed agent output is a representation the signer cannot truthfully make. Level 3 is compatible if and only if the three conditions of Section 4.2 demonstrably hold, and the sign-off readiness assessment of Section 8 operationalises that demonstration.

5. Why the Three Conditions Are Necessary

Each condition answers a documented property of the technology or of human oversight. This section states the property, cites its evidence base, and shows why the corresponding condition is the minimal defence. It closes with three failure archetypes, drawn from the author's development practice and consistent with the cited literature, that illustrate how the properties manifest in actuarial work specifically. Figure 2 summarises the derivation: each documented property of the technology or of human oversight generates one condition, and each condition resolves into a specific engineering defence.

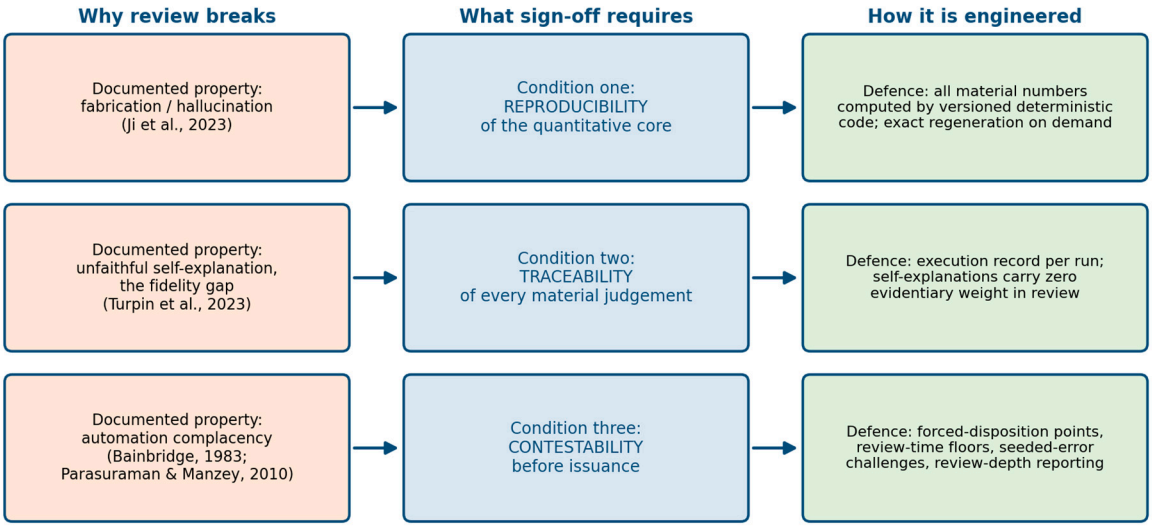

Figure 2. Derivation of the three conditions: documented properties, the conditions they necessitate, and the defences that implement them.

5.1. Fabrication and the Case for Reproducibility

Large language models generate fluent content that is not grounded in their inputs, a phenomenon extensively documented in the natural language generation literature under the heading of hallucination (Ji et al., 2023). The property is not an occasional defect but a structural feature of systems optimised to produce plausible continuations, and recent analyses locate its origin in the statistical objectives and evaluation incentives under which such models are trained (Kalai et al., 2025): plausibility and truth coincide often enough to be useful and diverge often enough to be dangerous. For actuarial work the implication is direct. Any quantity that a language model produces through generation, rather than through invocation of deterministic code, carries an irreducible risk of confident fabrication, and the risk is invisible precisely because fabricated numbers are formatted, contextualised and caveated exactly as computed numbers are. Condition one is the minimal defence: if every material number must regenerate exactly from identified inputs through versioned code, fabrication in the quantitative core is not merely detectable but architecturally excluded, and the residual fabrication risk is confined to the narrative layer, where conditions two and three address it.

 doi:10.20944/preprints202607.2097.v1

5.2. Unfaithful Self-Explanation and the Case for Traceability

The canonical review act, asking the delegate to explain its reasoning, presupposes that the explanation is causally connected to the reasoning. For large language models that presupposition is empirically false: models produce plausible explanations of their own outputs that demonstrably do not reflect the factors that actually drove those outputs (Turpin, Michael, Perez, & Bowman, 2023). This paper terms the resulting phenomenon, as it appears in agentic technical work, the fidelity gap: the distance between the rationale an agent states for a judgement and the process that generated the judgement, with the stated rationale optimised for plausibility rather than fidelity. The fidelity gap is the single most dangerous property of agentic actuarial systems for the accountability question, because it attacks the reviewing actuary at the precise point where professional training tells her she is doing her job. Interrogating the delegate's reasoning returns artifacts of the language layer, not evidence about the analysis, and subsequent work on reasoning-trace faithfulness reports the same dissociation even for models that expose extended chains of thought (Anthropic, 2025).

The practical rule this supports is stark: the agent's self-explanations must be treated as having zero evidentiary weight, and review must run exclusively on the execution record. Condition two follows as the minimal defence. It deliberately demands less than faithful explainability, which current systems cannot supply, and more than documentation, which describes the system rather than the run: it demands that every material judgement be locatable at an identifiable locus in the recorded workflow, so that the judgement can be evaluated on its merits. Traceability infrastructure is therefore not a documentation nicety; it is the only channel through which review of the judgement layer can be conducted at all.

5.3. Automation Complacency and the Case for Contestability

The third property belongs not to the technology but to its human overseers. The human-factors literature has documented for four decades that human monitoring of reliable automation degrades: attention migrates away from verification as the automation accumulates a record of success, and the operator's role collapses from active checking to passive ratification, precisely the dynamic Bainbridge (1983) identified as an irony of automation, and which subsequent research has integrated under the heading of automation complacency and automation bias (Parasuraman & Manzey, 2010), with experimental work documenting errors of omission and commission when human decision-makers defer to automated aids (Skitka et al., 1999). Two features of agentic actuarial systems make them an unusually strong complacency generator. Their output is fluent, which, as Section 2.3 argued, reads as competence to a reviewer trained on human work. And they are mostly right, which is what a well-built system delivers; every uneventful review weakens the behavioural case for the next one.

The implication is that deployment governance cannot rely on reviewer virtue, however conscientious the individual. Condition three must be engineered as a forcing structure: judgement points that require explicit human disposition before the report can be issued, review-time floors treated as control requirements, periodic seeded-error challenges to verify that review remains substantive, and management information that reports review depth alongside production volume. This also grounds the paper's central aphorism: the accountability boundary does not break where the technology fails, since failure tightens review; it breaks where the technology succeeds so smoothly that review quietly degrades into ratification.

5.4. Three Failure Archetypes in Actuarial Work

The general properties above manifest in actuarial work in recognisable patterns. Three archetypes deserve naming, because naming them is what allows a reviewing actuary to look for them. They are stated here as characteristic failure modes of language-model-based actuarial systems, consistent with the literature cited above and familiar from the author's own development practice; establishing their frequency in deployed systems is a task for the empirical protocol of Section 7.

The first archetype is confident interpolation: the interpretive layer characterises results that the quantitative core did not produce, for example describing tail behaviour that was never tested, in prose indistinguishable from grounded commentary. This is Section 5.1's fabrication property ex-

 doi:10.20944/preprints202607.2097.v1

pressed at the narrative layer. The second is silent scope reduction: where an analysis fails or data is insufficient, the system's default generation behaviour is to omit the analysis from the report without flagging the omission, producing a document whose apparent completeness is an illusion. The third is unearned reassurance: summary language drifts toward adequacy-affirming formulations stronger than the computed results warrant, a domain-specific expression of the tendency of generative systems toward agreeable, expected output, documented in the alignment literature as sycophancy (Sharma et al., 2023). None of these archetypes is a bug in the conventional sense. Each is the default behaviour of a fluent generative system doing what it was optimised to do, namely produce plausible, coherent, well-formed text. Honesty, in agent-produced actuarial work, is not a property that degrades under fault conditions; it is a property that must be engineered in against the grain of the underlying technology, and maintained under regression testing like any other requirement. The next section specifies the architecture that does the engineering.

6. A Reference Architecture for Accountable Agentic Actuarial Systems

6.1. The Separation Principle

The load-bearing design decision is the strict separation of the quantitative core from the linguistic layer. All reserve mathematics is performed by conventional, versioned, deterministic code that the agent invokes: development factor estimation and chain-ladder projection in the tradition formalised by Mack (1993), Bornhuetter-Ferguson projections combining a priori expected losses with emergence to date (Bornhuetter & Ferguson, 1972), expected loss ratio indications, and distributional reserve adequacy testing. The language model never computes a number that matters. Its roles are confined to orchestration, deciding which analyses to run and in what sequence; parameter proposal, suggesting factor selections and a priori assumptions which are then executed by the deterministic core; and interpretation, drafting the narrative review of the computed results. Every number in the final product is generated by the deterministic core and injected into the narrative by the application layer, never recalled or restated by the model. This division follows the tool-use pattern now standard in the technical literature, in which language models delegate computation to external programs rather than generating results token by token (Gao et al., 2023; Schick et al., 2023). Figure 3 depicts the resulting architecture and its information flows.

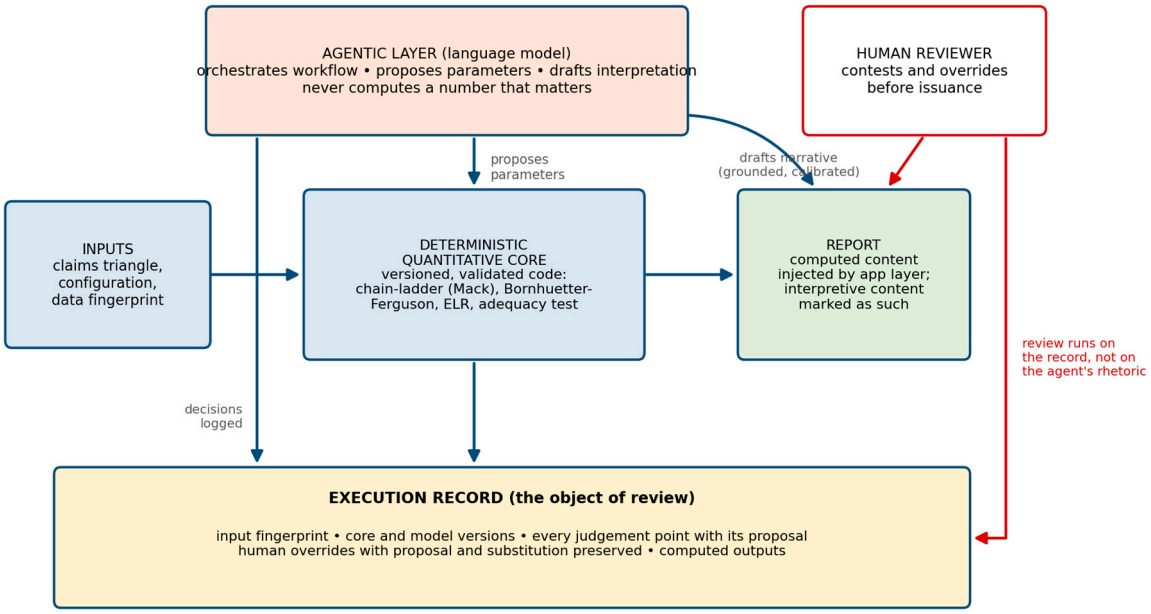

Figure 3. Reference architecture for an accountable agentic actuarial system: the separation principle, the execution record as the object of review, and the reviewer's contest-and-override path.

 doi:10.20944/preprints202607.2097.v1

This separation is the architectural expression of condition one, and its consequence is a clean partition of the review problem. The computed layer, once validated as software, needs no per-run review: given the same inputs and parameters it returns identical results, and an independent reviewer with the data and the parameter log can regenerate every figure. The judgement layer, by contrast, is not reproducible even in principle, because a stochastic language system proposing parameters in separate sessions will not propose identically; it can never be reviewed by re-running, only the way one reviews a human's judgements, on their merits, one at a time. Any deployment whose review protocol does not allocate human attention specifically and sufficiently to the judgement points is reviewing the wrong layer.

6.2. Traceability Instrumentation

In service of condition two, every run must produce an execution record comprising the input data fingerprint, the versions of the quantitative core and of the model, the full sequence of agent decisions with the proposed parameters at each judgement point, any human overrides applied with both the proposal and the substitution preserved, and the computed outputs. The report itself should distinguish typographically between computed content and interpretive content, so that a reviewer always knows whether a given sentence is arithmetic or rhetoric. The execution record, not the agent's narrative, is the object of review.

6.3. Honesty Engineering Against the Three Archetypes

Each archetype of Section 5.4 admits a specific architectural counter. Confident interpolation is countered by hard grounding: every claim in the narrative layer that references a quantitative result is validated against the computed result set before the report can render, and claims without a computed referent are blocked. Silent scope reduction is countered by mandatory completeness manifests: the report must enumerate every analysis attempted and its status, so that omission is impossible without disclosure. Unearned reassurance is countered by calibrated language rules: summary vocabulary is bound to quantitative thresholds, so that the strength of adequacy language is a function of the computed adequacy result rather than of the generator's stylistic drift. The common structure of the three counters deserves emphasis: in each case the defence is not a better prompt but a constraint enforced by the application layer outside the model, which is the only place a constraint on a generative system can be relied upon to hold.

7. An Evaluation Protocol for the Three Conditions

The conditions are only useful if they can be tested against a deployed system. This section specifies a protocol of three probes, one per condition, that an actuarial function, an internal auditor, a supervisor or a researcher can execute against any agentic reserving system. The protocol is stated prescriptively; its empirical execution, against the author's own production systems and others, is identified as the immediate next stage of this research programme.

7.1. Probe One: Reproducibility

Select a set of evaluation triangles, synthetic or anonymised. Process each through the full workflow, then regenerate every figure in the resulting reports directly from the run records using the quantitative core alone, without the agent. The condition requires exact regeneration for every figure; a single non-regenerable material quantity fails the probe and indicates that the language layer is producing numbers. Separately, submit each triangle twice in independent sessions with the quantitative core version pinned, and compare the agent's parameter proposals at each judgement point. The expected result, on current technology, is high agreement on mechanical selections and divergence at genuinely discretionary points such as tail factors and method weighting; the probe's purpose is not to demand judgement-layer reproducibility, which is unattainable, but to map exactly where the discretion sits, because those loci are where per-run human review must concentrate.

 doi:10.20944/preprints202607.2097.v1

7.2. Probe Two: Traceability and the Fidelity Gap

From the finished reports, list every material judgement. For each, attempt to trace the judgement to its locus in the execution record: a prompt instruction, a configuration parameter, a data feature or a recorded agent decision. The proportion traced measures locational traceability, and the condition requires that it be complete for material judgements. Then, separately, compare the report's stated rationale for each judgement against the recorded decision sequence, and count the instances in which the rationale is plausible, professionally phrased and inconsistent with the recorded path, for example a factor selection rationalised by reference to a stability criterion that the record shows was never evaluated. That count estimates the system's fidelity gap. Given the evidence on unfaithful self-explanation (Turpin et al., 2023), a positive fidelity gap should be the working assumption, and a measured gap of any size confirms the protocol's corollary: review protocols must assign the agent's self-explanations zero evidentiary weight.

7.3. Probe Three: Contestability

Contestability is probed mechanically and behaviourally. Mechanically: apply a human override at each class of judgement point and verify that it propagates through recomputation into the final product, with both the proposal and the substitution preserved in the record; this is precisely the audit trail a supervisor should ask to see. Behaviourally: log review time and review depth per run over the operating period, and test for the decline that the automation-complacency literature predicts (Parasuraman & Manzey, 2010), for example by comparing early and late review sessions or by seeding deliberate errors at known judgement points and measuring detection rates over time. A declining review-time series without a corresponding, documented improvement in system quality is presumptive evidence that contestability is degrading; seeded errors that pass review undetected are conclusive evidence. The behavioural half of this probe is the empirically novel one, and executing it across practising reviewers, rather than relying on the complacency literature's generic findings, is the highest-value experiment this framework makes possible.

8. The Sign-Off Readiness Framework

8.1. Purpose and Form

The framework distils the conceptual apparatus of Sections 4 to 6 and the protocol of Section 7 into an assessment that three audiences can apply: an appointed or responsible actuary deciding whether to rely on an agentic system, an audit committee or board interrogating management's claims about one, and a supervisor examining a firm that deploys one. The assessment proceeds in two stages: classification, then condition testing.

8.2. Stage One: Classify the Actual Delegation Level

The assessor first determines the system's operating level on the Section 4.3 hierarchy, as deployed rather than as documented. The diagnostic questions are behavioural. Does any judgement made by the system reach the final product without explicit human disposition? If yes, the system is at Level 3 or above regardless of what the governance documentation claims. Does any system output enter booking, filing or disclosure without item-level review of that output? If yes, the system is operating at Level 4 for that output, whatever the policy says. The complacency mechanism of Section 5.3 predicts that the most common misclassification will be claimed Level 2 masking actual Level 3: judgement points are nominally surfaced for human decision, but surfaced in such volume, or with such strong default proposals, that the human disposition is a click-through. Volume of surfaced decisions per unit of allocated review time is the tell-tale metric.

8.3. Stage Two: Test the Three Conditions

For systems at Level 3, the assessment then tests each condition using the Section 7 protocol. The diagnostic questions below are phrased so that an assessor can put them directly to management, a vendor or an internal team, together with the evidence each answer requires.

 doi:10.20944/preprints202607.2097.v1

Reproducibility:

Can every figure in a sampled report be regenerated exactly from the run record, without the agent's involvement, on demand and in the assessor's presence?

Is there architectural attestation that no material quantity is produced by the language layer, and does inspection of a run record corroborate it?

Are the quantitative core and its version history under conventional software validation and change control?

A team unable to demonstrate regeneration on demand fails the condition, whatever the documentation asserts.

Traceability:

For a sample of material judgements in a finished report, can each be traced to its locus in the execution record: a prompt instruction, a configuration parameter, a data feature or a recorded agent decision?

When the report's stated rationales are compared against the recorded decision sequence, is the fidelity gap measured at all, and how large is it?

Does a written review protocol confirm that agent self-explanations carry no evidentiary weight and that review runs on the execution record alone?

Contestability:

Does an override at each class of judgement point demonstrably propagate through recomputation into the final product, with both the proposal and the substitution preserved in the record?

Are there forced-disposition judgement points, review-time floors or equivalent controls that prevent review from collapsing into click-through, and are they monitored?

Do seeded-error challenges verify periodically that review remains substantive, and does management information report review depth alongside production volume over time?

A declining review-time series without a documented quality justification is presumptive evidence of degradation.

The framework's pass rule is conjunctive: reliance consistent with a statutory signature requires all three conditions evidenced, at the classified level, for the specific opinion in question. Partial satisfaction supports lower-level uses, drafting, exploration, challenge of human work, but not sign-off reliance. Figure 4 assembles the two stages and the pass rule into a single decision flow.

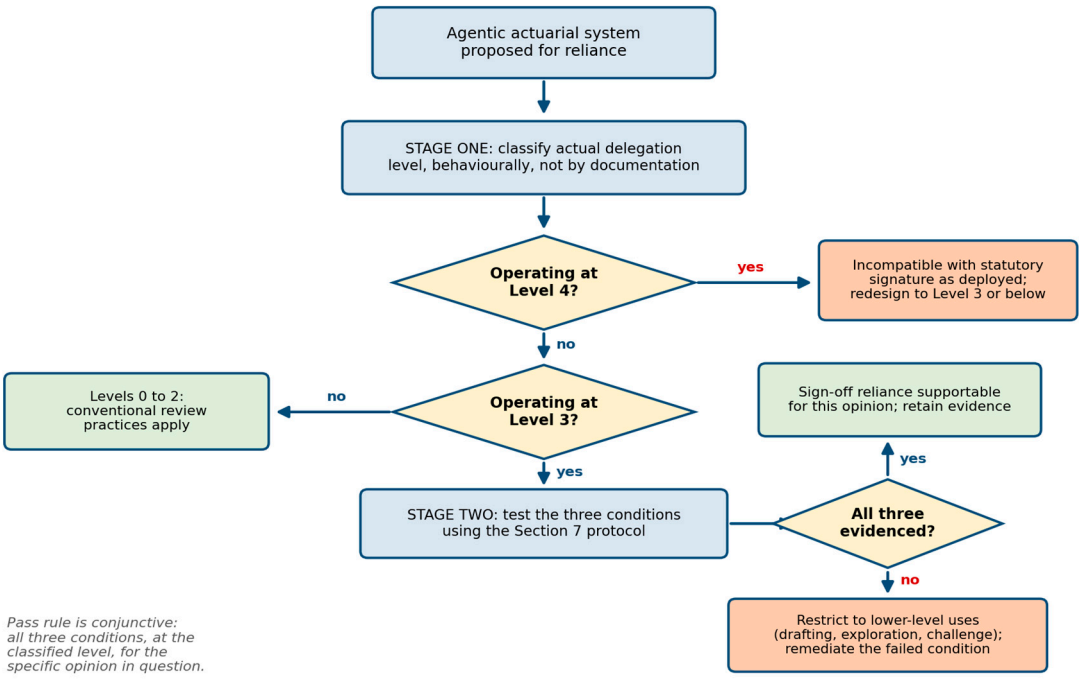

Figure 4. The sign-off readiness assessment as a decision flow: behavioural classification, condition testing under the Section 7 protocol, and the conjunctive pass rule.

 doi:10.20944/preprints202607.2097.v1

8.4. What the Signature Comes to Mean

Under the framework, the signing actuary's representation in respect of agent-produced work is precise: I have verified that the quantitative core reproduces; I have reviewed the material judgements at their traced loci on their merits, disregarding the system's rhetoric; and I retained and exercised the practical ability to contest before issuance. That representation is demanding but makeable, which is the point. The alternative representations on offer are either not demanding, the system is well governed in general, or not makeable, I have reviewed this work as I would a human's. The framework's contribution is to give the profession a representation that is both.

9. Discussion

9.1. Implications for Supervisors

For supervisors, the analysis suggests a shift in examination focus from program documentation to delegation reality. The instruments now being deployed, model bulletins, evaluation tools, governance guidance, generate paper that is necessary but examinable mostly at the institutional layer. The two questions this paper equips examiners to add are individually pointed: at what level of the delegation hierarchy is each material actuarial system actually operating, evidenced behaviourally; and can the responsible actuary demonstrate the three conditions for any system relied upon in a statutory opinion? Both questions are answerable from artifacts the firm should already possess, run records, review logs, override histories, and their absence is itself the finding. The direction of travel is favourable: the human oversight expectations articulated in the international application paper on artificial intelligence supervision (International Association of Insurance Supervisors, 2025) provide a natural docking point for delegation-level classification. For European supervision specifically, the analysis implies that human oversight requirements of the kind the Artificial Intelligence Act imposes on high-risk systems (European Parliament and Council, 2024), whose interaction with actuarial responsibilities the European profession has begun to map (Actuarial Association of Europe, 2025), are underspecified for statutory actuarial work unless they are interpreted to require contestability in the engineered sense of Section 5.3, not merely the formal presence of a human in the loop.

9.2. Implications for Professional Bodies

The reliance standards of the international actuarial profession require amendment, not because their principles are wrong but because their tacit assumptions, interrogability, error legibility, normative alignment, fail for artificial delegates as shown in Sections 3.2 and 5. The three conditions offer a drafting skeleton: reliance on system-produced work could be made permissible where the actuary has evidence of reproducibility, traceability and contestability proportionate to materiality, with the agent's self-explanations explicitly excluded from the evidence base. The delegation hierarchy offers standard-setters a scoping device: standards can attach obligations to levels rather than to technologies, which future-proofs the drafting against the next architecture. There is also a disciplinary corollary the profession will eventually face: if a signature is issued over Level 4 output, the defect is not negligence in review but the absence of review, and experience from other professions suggests the distinction matters. In auditing, disciplinary bodies have consistently treated opinions issued without the underlying audit work having been performed as a different and graver category of misconduct than opinions resting on work that was performed negligently, and there is little reason to expect actuarial tribunals to reason differently.

9.3. The Capacity-Constrained Market Dilemma

The accountability question is most acute where agentic tools are most valuable. In many African insurance markets the number of qualified actuaries is small relative to the number of licensed insurers, statutory valuation requirements are expanding under modernised insurance legislation, and the economics of agentic tooling are correspondingly compelling: a Level 3 system can extend one qualified actuary's effective capacity across engagements that would otherwise be unserved. The temptation, for firms and arguably for regulators facing compliance bottlenecks, is to let capacity pressure

 doi:10.20944/preprints202607.2097.v1

push deployments quietly toward Level 4. The analysis here suggests the opposite policy conclusion. Precisely because the reviewing layer is thin, capacity-constrained markets have the least redundancy to absorb the silent failure modes analysed in Section 5, and the strongest interest in mandating the three conditions as licensing requirements for actuarial artificial intelligence systems from the outset.

There is a genuine opportunity in sequencing. Markets that are now drafting their first artificial intelligence guidance for insurance, as several African supervisors are under national artificial intelligence strategies, can write delegation-level classification and condition testing into supervisory expectations before legacy deployments accumulate, an institutional leapfrog analogous to the mobile-money trajectory in payments. The author's related work on operationalising the insurance dimension of Ghana's national artificial intelligence strategy develops this supervisory design space in detail (Botchey, 2026).

9.4. Trust, and the Political Economy of Fluency

Finally, the analysis connects to the empirical adoption literature. The author's doctoral research on digital insurance adoption in Ghana found vendor trust to be a stronger driver of adoption than perceived usefulness (Botchey, 2025), a pattern consistent with a broader institutional argument that in environments of uncertainty, adoption follows trust signals rather than verified performance. Fluency is the most powerful trust signal an agentic system emits, and Sections 2.3 and 5 argue it is also among the least informative. The commercial equilibrium this implies is troubling: market rewards accrue to fluency, while the properties that make sign-off meaningful, reproducibility, traceability, contestability, are costly, invisible in a demonstration, and easily claimed without being engineered.

This is a classic market for lemons in verification properties (Akerlof, 1970), and it will not correct itself. The correction has to come from the demand side, from the professionals whose signatures are at stake and the supervisors who stand behind them, equipped with tests that pierce the fluency. Supplying such a test is what this paper has attempted.

10. Limitations and Future Research

Four limitations bound the claims. First, the paper is conceptual: the three conditions are derived from documented general properties of the technology and of human oversight, and from disclosed practitioner experience, but they have not yet been validated by systematic empirical application of the Section 7 protocol; that application, against the author's own production systems and against third-party systems, is the immediate next stage of this research programme. Second, the illustrative domain is general insurance reserving; life valuation, capital modelling and pricing embed different judgement structures and may strain the conditions differently, and the framework's portability across those domains is asserted, not demonstrated. Third, the behavioural argument for contestability rests on the general automation-complacency literature rather than on studies of actuarial reviewers specifically; a controlled study of review degradation among practising actuaries, which the seeded-error design of Section 7.3 makes straightforward to construct, would materially strengthen or qualify the claim. Fourth, the technology is moving: architectures that expose faithful reasoning traces, if they materialise, would narrow the fidelity gap and soften the strict evidentiary exclusion of self-explanations, though the framework anticipates this by making the exclusion conditional on infidelity rather than axiomatic. Future work should also include standard-setting engagement with the professional bodies whose reliance provisions this paper has argued are due for amendment.

11. Conclusion

The actuarial signature is one of the oldest accountability technologies in financial regulation, and it is about to be tested by the newest. This paper has argued that the test is subtler than the prevailing governance conversation suggests. The danger is not that artificial agents will do actuarial work badly; increasingly, they will do it well. The danger is the one the documented properties of the technology and of human oversight jointly predict: the accountability boundary does not break where the technology fails; it breaks where the technology succeeds so smoothly that review quietly

 doi:10.20944/preprints202607.2097.v1

degrades into ratification, so that the signature drifts from a representation of verified judgement to a ceremonial countersignature on machine output, without anyone deciding that it should.

Against that drift the paper has set three conditions, reproducibility, traceability and contestability, each derived from a documented property rather than asserted as a principle; a five-level delegation vocabulary; a reference architecture that makes the conditions buildable; and an evaluation protocol and readiness assessment that make them demandable. The signature can survive the agentic era, but only as a representation that has been redefined with precision and then actually earned, one opinion at a time. Deciding to require that is work for supervisors and professional bodies. Showing that it is specifiable, and what it costs, has been the work of this paper; showing empirically where deployed systems stand against it is the work of the next.

References

Actuarial Association of Europe. (2025). Navigating Europe's AI Act: Insights for actuaries and the insurance sector. (Discussion paper).

Actuarial Standards Board. (2019). Actuarial Standard of Practice No. 56: Modeling.

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. Quarterly Journal of Economics, 84(3), 488-500.

American Academy of Actuaries. (2026). Actuarial algorithmic accountability: Setting ethical standards for AI. https://actuary.org

Anthropic. (2025). Reasoning models don't always say what they think [Research paper].

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

Bellis, C., Lyon, R., Klugman, S., & Shepherd, J. (Eds.). (2010). Understanding actuarial management: The actuarial control cycle (2nd ed.). Institute of Actuaries of Australia.

Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency. (2011). Supervisory guidance on model risk management (SR Letter 11-7).

Bornhuetter, R. L., & Ferguson, R. E. (1972). The actuary and IBNR. Proceedings of the Casualty Actuarial Society, 59, 181-195.

Botchey, O. K. (2025). Impact of digitalization on the Ghanaian insurance industry: Consumer behaviour and industry transformation (Doctoral dissertation). SBS Swiss Business School.

Botchey, O. K. (2026). Operationalising Pillar 6 of Ghana's National Artificial Intelligence Strategy in the insurance sector (Working paper). SSRN. https://ssrn.com/abstract=6768198

England, P. D., & Verrall, R. J. (2002). Stochastic claims reserving in general insurance. British Actuarial Journal, 8(3), 443-518.

European Parliament and Council. (2009). Directive 2009/138/EC on the taking-up and pursuit of the business of insurance and reinsurance (Solvency II). Official Journal of the European Union.

European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

Financial Reporting Council. (2023). Technical Actuarial Standard 100: General actuarial standards.

Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., & Neubig, G. (2023). PAL: Program-aided language models. In Proceedings of the 40th International Conference on Machine Learning (PMLR 202).

Institute and Faculty of Actuaries. (n.d.). The agentic model office: AI agents and the future of insurance organisations (THINK Thought Leadership Series, Issue 13). https://actuaries.org.uk

International Actuarial Association. (2025). Artificial intelligence governance framework (Artificial Intelligence Task Force paper). https://actuaries.org

International Actuarial Association. (2018). ISAP 1: General actuarial practice.

International Association of Insurance Supervisors. (2025). Application paper on the supervision of artificial intelligence. https://iais.org

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate. arXiv:2509.04664.

 doi:10.20944/preprints202607.2097.v1

Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. ASTIN Bulletin, 23(2), 213-225.

National Association of Insurance Commissioners. (2023). Model bulletin: Use of artificial intelligence systems by insurers.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381-410.

Prince, A. E. R., & Schwarcz, D. (2020). Proxy discrimination in the age of artificial intelligence and big data. Iowa Law Review, 105, 1257-1318.

Republic of Ghana. (2021). Insurance Act, 2021 (Act 1061).

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems, 36.

Sharma, M., et al. (2023). Towards understanding sycophancy in language models. arXiv:2310.13548.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? International Journal of Human-Computer Studies, 51(5), 991-1006.

Society of Actuaries Research Institute & Casualty Actuarial Society. (2026). 19th annual survey of emerging risks.

Swiss Confederation. (2004). Insurance Supervision Act of 17 December 2004 (as amended).

Swiss Financial Market Supervisory Authority FINMA. (2024). Guidance 08/2024: Governance and risk management when using artificial intelligence.

The Actuary. (2026, June 25). Draw the line: A governance framework for AI agents in insurance and finance. https://theactuary.com

Thompson, D. F. (1980). Moral responsibility of public officials: The problem of many hands. American Political Science Review, 74(4), 905-916.

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. Advances in Neural Information Processing Systems, 36.

Wüthrich, M. V., & Merz, M. (2023). Statistical foundations of actuarial learning and its applications. Springer.

Zhu, Q. (2026). Insurance of agentic AI. arXiv:2606.05449.