HUMAN CENTRED AI LAB
AI DECODED · UTS DATA SCIENCE INSTITUTE

2026

AI GOVERNANCE · DIRECTOR LIABILITY

Beyond *Probabilistic* Governance

Why Continuous Auditing Is the Only Legal Defence for Agentic AI

Sam Banerjee
AI Decoded · Researcher, University of Technology Sydney

Professor Jianlong Zhou
Associate Professor · University of Technology Sydney

EXECUTIVE SUMMARY

Your board approved an AI governance policy. Your risk committee reviewed the model accuracy reports. Your legal team signed off on the compliance framework. And your AI agent just executed a transaction that breached your risk appetite — autonomously, at speed, without a human in the loop.

You have a problem. And under section 180 of the Australian Corporations Act, the argument that it is a personal one will be difficult to resist.

Australian boards are operating under a governance assumption that was already dangerous and is now legally obsolete: that a high model accuracy score, or a point-in-time audit, constitutes a defensible compliance posture for Agentic AI. It does not. Probabilistic Variance — the inherent uncertainty in every AI model — is no longer a technical footnote. It is an unmitigated legal liability.

This paper introduces the Deterministic Risk Architecture (DRA): an engineering approach that wraps the probabilistic AI model in a deterministic shell, grounded in the Causal Quantification Framework developed by Professor Jianlong Zhou. The result is a compliance architecture that does not detect failure. **It prevents it.**

Directors who act now can move from hoping their AI systems behave within policy, to proving — mathematically and in real time — that they cannot do otherwise.

> **Continuous Auditing is not an upgrade to your current governance framework. It is the replacement for it.**

95%
Accuracy score boards treat as governance safe harbour

1/20
Executions that will fail — now execute without human review

230
FS AI RMF control objectives — impossible to satisfy manually

01

THE PROBLEM

The "Probabilistic Variance" Trap

Why a 95% accuracy score is a governance liability, not a safe harbour

Your AI agent made a decision this morning that your board never approved. Not because your governance failed. Because your governance was never designed to stop it.

Right now, across Australian boardrooms, Agentic AI systems are executing consequential actions — initiating transactions, finalising agreements, allocating resources, rejecting applications — at a volume and velocity no human oversight process can match. Boards have responded the way boards respond to technical complexity: policies approved, audits commissioned, performance reports tabled and noted. The right things, done correctly. None of it is enough.

> ***What was a manageable error rate in a supervised model is, in an agentic context, a guaranteed liability incubator.***
>
> BEYOND PROBABILISTIC GOVERNANCE · 2026

The Accuracy Illusion

When a board receives a model performance report showing 95% accuracy, the inference is reassurance. That inference is wrong. In an agentic context, it is dangerous. 95% accuracy means the system produces an incorrect or out-of-policy output once in every twenty interactions. In a decision-support model, the human catches it. Agentic AI removes the human. The agent does not recommend. It executes.

The Compliance Architecture That No Longer Exists

ETSI TS 104 008 V1.1.1 — CABCA — establishes Continuous Auditing as the recognised technical baseline for AI governance. A quarterly audit of an agentic system executing thousands of decisions per day is not governance. Under CABCA, it does not qualify as compliance. NIST AI 800-4 (2026) reinforces the global direction of travel. A governance architecture adequate in 2023 no longer reflects the recognised standard of care in 2026.

ETSI TS 104 008 V1.1.1

CABCA — Continuous Auditing Baseline

January 2026. Establishes Continuous Auditing as the recognised technical baseline. Makes point-in-time compliance legally inadequate.

NIST AI 800-4

Post-Deployment Monitoring Gaps

March 2026. Documents the absence of continuous oversight as a critical gap. Retrospective monitoring leaves organisations exposed.

FS AI RMF V1.0

Financial Services AI Risk Framework

February 2026. U.S. Treasury / Cyber Risk Institute. 230 discrete control objectives. Manual compliance is mathematically impossible.

CORPORATIONS ACT S.180

Director Duty of Care and Diligence

A conduct standard. A governance failure becomes a personal liability question before any damage is proven.

The Verification Gap

02

THE SHIFT

From Explainability to Enforceability

Why XAI is a flight recorder, not a safety system

The first wave of AI governance was built on a reasonable premise: if we can explain what the AI did, we can defend it. It solved the wrong problem.

Explainability is a forensic capability. It is the corporate equivalent of a black box flight recorder — extraordinarily useful for understanding what went wrong, entirely useless for preventing the crash. Enforceability operates upstream of the model's decision process. It asks: "Is this decision one the AI is authorised to make?" And it asks before execution — not after.

A faster alarm is not a lock on the door. Detection is not prevention.

BEYOND PROBABILISTIC GOVERNANCE · 2026

The Causal Quantification Framework

The Causal Quantification Framework developed by Professor Jianlong Zhou operates prospectively — it examines the real drivers of an AI system's decisions as they occur, not just the outcomes. While a standard audit might reveal that a model rejects older candidates more often, the Framework can uncover whether this is due to a direct age bias or a proxy — such as a "digital literacy score" acting as a stand-in for age.

Trustworthiness metrics are the framework's actionable outputs — indicators that light up when decision-making drifts towards risky or unauthorised territory, generated continuously in real time. Like a speedometer that updates every second, not just at scheduled service intervals.

> **WHY MEASUREMENT ALONE IS NOT ENOUGH**
>
> A governance system that measures deviation and reports it remains a Post-Hoc Observation system — it has reduced the detection lag from weeks to milliseconds, but detection is not prevention. Closing the Verification Gap requires that measurements be connected directly to an **enforcement mechanism** — one that acts before the agent's decision becomes an agent's action.

03

THE SOLUTION

Deterministic Risk Architecture

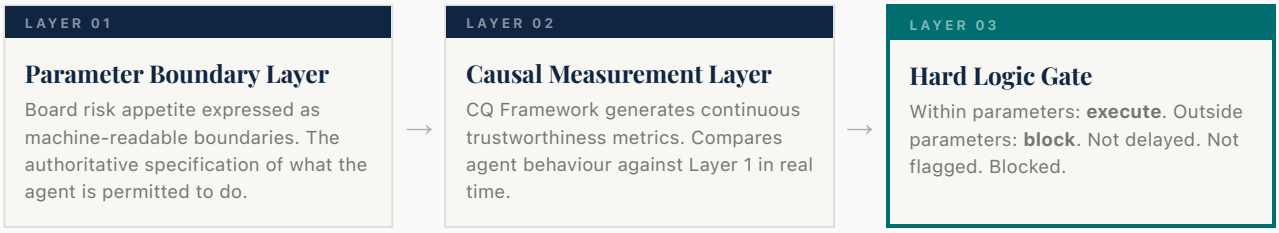
The engineering solution to the governance problem

Every significant safety-critical industry solved the problem of unreliable components the same way. They built deterministic systems around them. The DRA is not a monitoring dashboard. It is an enforcement layer.

Think of it as the difference between a speed camera and a speed limiter. A speed camera observes violations after the fact. A speed limiter makes exceeding the designated speed mechanically impossible. Current AI governance is a speed camera. The DRA is a speed limiter.

FIGURE 1 — THREE-LAYER ENFORCEMENT ARCHITECTURE

The model remains probabilistic. The governance outcome is deterministic.



The board's risk appetite is not a policy aspiration. It is a hard boundary the agent is architecturally incapable of crossing.

BEYOND PROBABILISTIC GOVERNANCE · 2026

A board that has deployed DRA can produce, on demand, the formal specification of its agent parameter boundaries, demonstrate the enforcement mechanism that makes deviation structurally impossible, and provide a continuous audit record documenting every governance decision. That is not a compliance report. It is computational proof.

04

THE IMPERATIVE

Why Boards Must Act Now

Five concrete actions for Directors — and why they are personal

At some point in the next few years, an Australian court will determine what reasonable diligence means for a Director whose organisation deployed Agentic AI. The governance decisions that will determine its outcome are being made right now.

Section 180 of the Corporations Act is a conduct standard — the question a court will ask is not whether something went wrong, but whether the Director's conduct met the required standard of care. A governance failure can become a question of personal liability before any damage is proven.

Hope is not a governance framework. And under s.180, it is unlikely to function as one in court.

BEYOND PROBABILISTIC GOVERNANCE · 2026

Five Imperatives for Directors

Bring these questions to your next risk committee meeting

I **Diagnose your architecture.** Establish whether your current AI governance is built on Post-Hoc Observation or Deterministic Constraint. If your primary evidence of AI compliance is model accuracy reports and audit logs, the Verification Gap is open.

II **Demand the enforcement specification.** Require technology and risk leadership to specify, in writing, the enforcement mechanism that prevents your Agentic AI systems from operating outside board-approved parameters. If it cannot be produced, it does not exist.

III **Measure the distance.** Assess your organisation's exposure against ETSI CABCA and NIST AI 800-4 and determine the distance between your current governance architecture and the Continuous Auditing baseline.

IV **Own the governance decision.** Treat the transition to Continuous Auditing as a board-level priority, not a technology project delegated to management. The exposure under s.180 is personal. The governance decision must be personal, too.

V **Require computational proof.** Engage with the Deterministic Risk Architecture as the engineering answer to the question your board should already be asking: how do we *prove*, not assert, that our AI systems cannot breach our risk appetite?

The decision is now a board decision.

The history of corporate governance is a history of boards learning, often through painful experience, that the governance frameworks adequate for one technological era are insufficient for the next. The financial crisis demonstrated that risk models boards had treated as reliable were probabilistic constructs whose failure modes were neither understood nor governed.

Agentic AI is the next inflection point. The probabilistic models at its core carry the same structural uncertainty — with the addition of autonomous execution capability that removes the human buffer between model error and organisational consequence.

The boards that govern this moment well will not be the ones that deployed the most sophisticated AI. They will be the ones that understood, early enough to act, that deploying Agentic AI without Continuous Auditing was not a governance gap to be managed. It was a liability to be eliminated.

The architecture exists. The legal imperative is established. The regulatory baseline is set.