aveni

THE MILLS REVIEW

Six steps to Mills readiness

The Mills Review changes the industry's inherent risk profile, not just the control environment around it. This paper sets out what closing that gap requires: a board level view of why the exposure is already live, and the six steps that move a firm from policy on file to evidenced, operational assurance.

THE SIX QUESTIONS AT A GLANCE

STEPS 01 TO 06

01 Have we classified every AI use case by autonomy level and risk tier?

RISK ASSESSMENT FRAMEWORK

02 Is our policy written as controls an agent can actually be tested against?

ENFORCEABLE CONTROLS

03 Can we assure every interaction continuously, instead of sampling once a year?

CONTINUOUS ASSURANCE

04 Are the models doing that assurance live, layered and explainable?

MODEL SELECTION

05 Could we evidence any single decision on demand, without weeks of assembly?

EVIDENCE BACKBONE

06 Does one standard of assurance cover humans, copilots and agents alike?

UNIFIED OVERSIGHT

Continuous assurance | Agent governance | Evidence by default

CONTEXT · WHY NOW

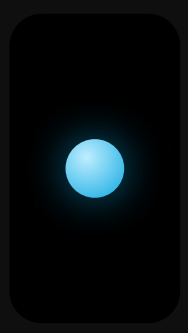
Why now

Every major regulatory shift has changed the infrastructure of financial services. Basel changed capital. GDPR changed data. Consumer Duty changed customer outcomes. The Mills Review is different in kind: it is the regulator's acknowledgement that the industry's inherent risk profile is changing, not merely the control environment around it.

Inherent risk is increasing by orders of magnitude. Financial services governance assumes a human adviser makes one judgement at a time, within a bounded caseload, an error stays contained to their own book of work and is caught at the next review cycle. Agentic AI removes that constraint: a single agent applies the same judgement, correct or flawed, across an entire customer population simultaneously. This is a change of orders of magnitude in the population a single failure can reach, and in what the firm is liable for if it goes undetected. A firm that has not re-quantified its inherent risk for AI-enabled use cases is carrying exposure it has not measured.

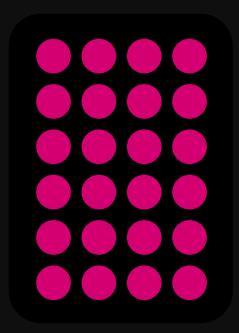
TODAY

A human adviser handles a bounded caseload. Judgement is applied one customer at a time. An error is contained to their own book of work, and caught on the next review cycle.



TOMORROW

One agent applies the same judgement across the entire eligible customer base, simultaneously. A flawed pattern doesn't wait for a review cycle, it's already live everywhere the agent operates.



The likelihood of failure is not currently knowable. Most AI guardrails were built to catch factual error, not the failure mode that matters most in regulated decisioning: a system that is subtly wrong while sounding entirely confident. There is no actuarial base rate to draw on, no loss history, no benchmark for how often a given class of agentic system should be expected to fail at scale. Firms citing a low likelihood of failure today are, in most cases, asserting confidence they don't yet have evidence to support.

The controls proposed to manage this are themselves new and unproven. Continuous monitoring, live supervisory data, agent-level accountability, an Agentic Supervisory Model at the regulator, each is sound in principle, but none has been tested at scale in a live regulated environment. A control framework, including a firm's own, should be treated as unproven until it has demonstrated, under real operating conditions, that it can catch a drifting or failing agent before harm reaches a material share of the customer base.

> "Larger exposure, unknown probability, and untested mitigation is precisely the profile that should trigger Board-level risk appetite discussion, not a second-line policy update."

What happens before the new framework is in place. The Review's recommendations will take time to work through, but risk exposure and appetite are increasing now, regardless of that timetable. Commercial pressure to deploy agentic AI is intensifying as firms look to capture the productivity gains it promises, while the control model most firms hold has not caught up with the one they will need. Silence from Risk is not neutral in this gap - it is an unquantified, unapproved risk position. Boards should expect a number and a mitigation plan on the table now, not a policy document held on file until the framework catches up. Four things move in the interim, whether or not a firm has acted:

01

Risk appetite and exposure shift

The population a single failure can reach has already changed, whatever a firm's risk statement says.

02

SM&CR accountability grows

Senior Managers carry personal accountability for agentic activity today, not from a future date.

03

Capital and PI positions are exposed

Insurance and capital positions were priced against a human-scale risk profile that no longer holds.

04

Reputational and regulatory exposure rises

The FCA has stated its expectations publicly; current practice is read against that standard already.

01 Establish a risk assessment framework

Every AI use case, existing and planned, must be classified before controls, monitoring, or evidence requirements can be calibrated. This requires two things together: placing the use case on the Mills autonomy spectrum, and assigning it a risk tier.

LEVEL	ROLE	EXAMPLE
L1 Operator	Human uses AI as a tool	Summarises product terms
L2 Collaborator	Human and AI plan and act together	Compares products with user refinement
L3 Consultant	AI recommends, human decides	Builds a plan, seeks preferences
L4 Approver	AI prepares, human authorises	Prepares a transfer for approval
L5 Observer	AI acts within boundaries, human monitors	Executes actions, logs for review

Regulatory pressure concentrates at L4 and L5, where "human in the loop" stops being sufficient on its own, firms must state precisely what the person does, what they see, when they intervene, and how that challenge is recorded. Layered on top, each use case gets a risk tier by potential impact:

LOW RISK · ASSISTIVE — HIGH RISK · AUTONOMOUS

A — Low-risk assistive
Internal or low-impact support; no customer-facing action; no tool execution.

APPROVAL EXPECTATION
Standard pre-prod review · control checklist · proportionate monitoring

B — Customer-facing constrained
Customer-facing interactions within clear scripts or bounded actions; limited tool usage.

APPROVAL EXPECTATION
Formal approval · stress testing · defined policy controls · live assurance

C — Customer-affecting agentic
Meaningful customer impact; tool execution; judgement-like recommendations; potential vulnerability exposure.

APPROVAL EXPECTATION
Enhanced approval · deeper stress testing · explicit human approval · tight thresholds · escalated live review

D — High-risk autonomous
Material customer, regulatory or financial impact; broad permissions; complex workflows; high autonomy.

APPROVAL EXPECTATION
Executive-level approval · strongest restrictions · mandatory human gates · intensive monitoring · re-approval required

This determines *how much* control is required, not *whether* - even Tier A needs defined ownership and evidence, proportionately less than Tier D.

02 Translate policy into enforceable controls

Second-line policy must become part of the operating fabric - rules, thresholds, and test criteria an agent can be checked against, before and continuously after it reaches a customer. Explainability is too often treated as a model property ("is this model interpretable?") rather than an operational one ("can we explain this decision, to this person, right now?"). The second is what a supervisor will actually test, and it needs to be built into the workflow itself, not reconstructed after the fact.

Before any use case goes live, capture a structured approval record: use case, ownership and customer context; autonomy, tools and control boundaries (including kill-switch conditions); policy and testing requirements; live supervision requirements. This record is not filed away after sign-off, it is the reference point for how the use case is monitored in production.

RISK DOMAIN	WHAT SECOND LINE ASKS	TYPICAL FAILURE MODES	WHAT APPROVAL SHOULD DEFINE
01 Autonomy & authority	What can the agent do without human approval?	Unauthorised actions, excessive permissions, uncontrolled tool use, weak human gating.	Autonomy tier, approved & prohibited actions, human approval points, kill-switch conditions.
02 Customer and conduct risk	Could the interaction create unfair, unsuitable or harmful outcomes?	Poor treatment, weak disclosures, unsuitable prompts, weak vulnerability handling, scaled misconduct.	In-scope journeys, vulnerable-customer rules, conduct requirements, outcome tolerances.
03 Behaviour and response risk	Is the agent likely to respond safely and consistently?	Hallucination, misleading answers, inconsistency, unsafe wording, weak escalation.	Required guardrails, response boundaries, fallback logic, escalation triggers.
04 Tool and action execution risk	Can the agent safely use tools, data and downstream systems?	Incorrect tool invocation, unapproved actions, workflow breakdown, bad hand-offs.	Tool permissions, tool-specific controls, action limits, supervisory checkpoints.
05 Data, privacy and access risk	Is the agent accessing and using data appropriately?	Over-collection, weak minimisation, inappropriate retrieval, poor entitlements.	Data sources, entitlement rules, retrieval constraints, logging standards.
06 Explainability, evidence & traceability	Can the bank prove what happened and why?	Missing logs, weak lineage, inability to replay actions, weak justification.	Minimum evidence set, traceability rules, lineage capture, retention & replay standards.
07 Accountability and third-party dependency	Is ownership clear across the full agentic stack?	Blurred ownership, vendor opacity, unclear accountability, control gaps between teams.	Named owner, model/vendor accountability, RACI, dependency controls.
08 Change and lifecycle governance	What happens when the agent changes after launch?	Uncontrolled prompt changes, silent drift, unreviewed tool additions, degraded controls.	Material-change triggers, re-testing rules, re-approval thresholds, ongoing monitoring.

CONCENTRATION RISK

Control boundaries must also cover model and vendor concentration

The Review warns that shared reliance on a small number of model providers and hyperscalers creates correlated failure risk across the industry - a single upstream model issue can affect many firms at once. Approval records should therefore capture which models and vendors a use case depends on, whether the firm can access, query, and audit its own evidence independently of that vendor, and what happens to control and evidence if the vendor changes or fails. Evidence that lives entirely inside a third party's infrastructure is not evidence the firm can reliably stand behind.

03 Build continuous, full-coverage assurance

The most significant shift the Review points toward: from periodic compliance to operational assurance. An annual review of a system making thousands of decisions a day is closer to archaeology than oversight.

THE OLD WORLD		THE NEW WORLD
• Annual review	→	• Continuous monitoring
• Sampling		• Evidence generation
• Exception management		• Real-time intervention
• Board reporting		• Operational supervision

The Three Lines of Defence remain the right model, but each line's role changes: first line moves into live supervision - runtime thresholds, approval points, kill-switch authority; second line moves from periodic review to continuous control orchestration; third line stops reconstructing isolated cases and instead validates the control system itself, the validator of the validator.

1ST LINE

Agent operations and live supervision

The first line still owns risk, but what it owns fundamentally changes. First-line teams won't only manage business processes executed by people, they'll run mixed operations in which humans, copilots and autonomous agents all contribute to customer outcomes. That means first-line ownership must include:

- Runtime thresholds and permissions
- Defined autonomy boundaries
- Fallback routes and escalation paths
- Human approval points for sensitive actions
- Live monitoring of customer outcomes
- Intervention and kill-switch authority

In practice, first-line management becomes part operations leader, part digital supervisor and part control owner.

2ND LINE

Continuous control orchestration

The second line changes most. It moves from periodic review to dynamic control design, approval and continuous verification. The second line future state should include five core roles:

- **Policy translation** - convert policy requirements into machine-readable rules, thresholds and control tests
- **Approval governance** - approve agentic use cases through a consistent pre-production workflow
- **Live supervision oversight** - define what Tier 1 and Tier 2 monitoring must check in production
- **Change governance** - decide what changes require notification, re-testing or re-approval
- **Cross-risk integration** - bring together conduct, compliance, operational risk, model risk, privacy and customer outcome oversight

3RD LINE

Assurance-system validation

The third line remains as independent assurance, but its focus shifts upward. It's no longer sufficient to reconstruct isolated cases after the fact. Internal Audit must assess whether the bank's approval, monitoring, evidence and intervention architecture is reliable enough for a machine-operated environment. Key third-line questions become:

- Are autonomy tiers and approval gates consistently applied?
- Do monitoring controls cover the right risks and interactions?
- Is evidence complete, replayable and reliable?
- Are interventions timely and effective?
- Can management demonstrate control at Group level?
- Are second-line workflows, challenge and reporting genuinely independent and effective?

The third line becomes the validator of the validator.

04 Select models built for live, explainable assessment

Step 3 sets out the assurance *process* - who reviews what, and when. Step 4 addresses a separate question underneath it: whether the models doing that reviewing are actually fit for the job. A monitoring model that cannot assess every interaction as it happens, or cannot show its reasoning when challenged, is not fit for purpose regardless of how well it performs on average. Three criteria should govern model selection:

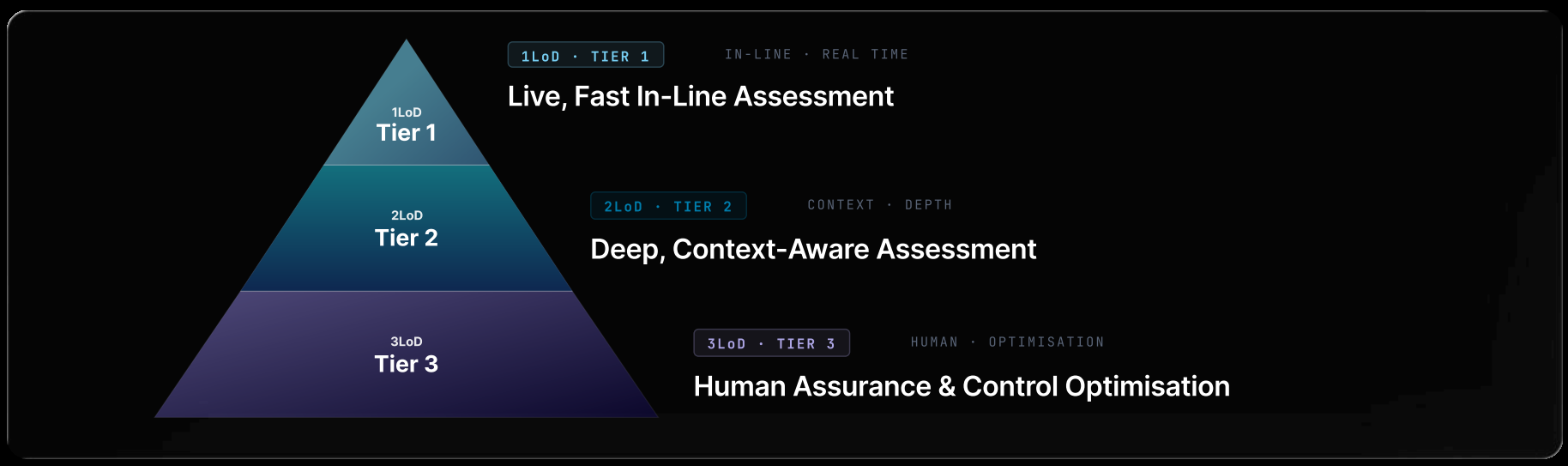
1. **Turn-by-turn, not batch.** The model must be capable of assessing every interaction live, as it happens, not on a delayed or periodic batch, since a flawed pattern in an agentic system is already live across the customer base from the moment of deployment.

2. **Depth matched to signal.** A single monitoring model rarely does both jobs well. Fast, lightweight models are better suited to constant, low-latency screening across every interaction; specialist, more contextually aware models are better suited to the smaller number of cases that screening flags as needing deeper interpretation.

3. **Explainable by construction.** The model must produce evidence and rationale for its own assessment, not just a pass or fail, that rationale is what a supervisor, an auditor, or a customer complaint will ultimately test. This is a property of how the model is built, not something that can be added afterward as a reporting layer.

General-purpose models are not disqualified by this, but they carry a burden of proof: a firm should be able to show why a given model's reasoning can be trusted and audited for this specific purpose, not simply that it performs well in general use.

Assessment, triage & human oversight



TIER 1

Live, fast in-line assessment

Small, specialist models operate live inside production latency budgets, assessing every agent turn in real time to identify, contain, or escalate risk before it reaches the customer - always-on, optimised for speed, precision and scale, without slowing the customer journey.

TIER 2

Deep, context-aware assessment

Specialist language models apply richer context, policies, guardrails and control frameworks to escalated cases with greater depth, judgement and explainability, producing evidence and rationale for risk, compliance and governance decision-making.

TIER 3

Human assurance & control optimisation

Human experts review complex or contested cases, validate that model-led controls operate within risk appetite, and feed learning back into Tier 1 and Tier 2 to improve performance and control design over time.

COMMAND CENTRE

Unified Assurance Command Centre

The single pane of glass across the full control stack - agent and human assessment in one customer-centred view of risk, insight and governance evidence for risk committees, boards and regulatory engagement.

05 Build the evidence and reporting backbone

Firms must be able to explain a decision - to a customer, to internal oversight, and to the regulator - on demand, not after weeks of manual assembly. Firms whose evidence lives in documents will be rebuilding it under supervisory pressure; firms whose evidence lives in systems will simply grant access.

Governance data should be reportable by default, so the Board, second line, and regulator draw from the same underlying evidence. That rests on three capabilities:

01

Traceability and lineage

End to end across prompts, models, and the actions taken on the back of them.

02

A common MI layer

One shared set of numbers for second line, senior management, Board, and Audit.

03

Replay capability

Reconstruct exactly what an agent saw and decided, for any single interaction.

06 Unify oversight across humans and AI

A single assurance model across the whole workforce: human, copilot, and agent, with consistent evidence standards, escalation paths, and accountability. Most firms are currently building AI governance as a bolt-on, producing two parallel regimes that will diverge and eventually conflict. Customers, Boards, and regulators will not accept differing evidence standards depending on who, or what, did the work.

In practice that means holding three things constant, whoever or whatever did the work:

01

One issue taxonomy

A single escalation model, regardless of the root cause behind an issue.

02

Comparable assurance

The same standard whether an interaction was human-led, copilot-assisted, or agent-led.

03

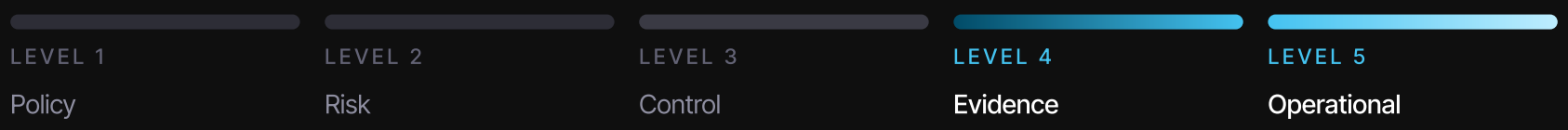
Enterprise visibility

Outcomes seen whole, rather than split across fragmented channel reporting.

Accountability has to be named, not assumed

Every agent should have a registered owner and be mapped explicitly to a Senior Manager's Statement of Responsibilities, so the same standing question can be asked of any channel, human or machine: **who owns this, what can it do, and who answers if it fails.**

Understand your AI Governance Maturity



LEVEL 1
Policy-driven
"We have an AI policy."

Governance exists as documentation. AI use is permitted or prohibited on paper; nobody can say with confidence what is actually happening.

LEVEL 2
Risk-driven
"We review AI."

AI systems pass through risk assessment before deployment. Governance is a gate at the start of the lifecycle, largely silent thereafter.

LEVEL 3
Control-driven
"We test AI."

Systems are tested before release and periodically re-tested. Controls exist, but evidence is point-in-time and ages quickly.

LEVEL 4
Evidence-driven
"We continuously monitor AI."

Governance produces a live evidence base. Performance, fairness and safety are observed in operation, not inferred from testing.

LEVEL 5
Operational
"Governance happens automatically."

Assurance is embedded in the infrastructure. Evidence is generated as a by-product of operation, intervention is real-time, and supervisory reporting draws directly from operational data.

Most firms sit at Level 1–2 today; regulatory direction points firmly toward 4–5. That gap, not model failure, is the defining AI risk firms carry into the next three years.

When oversight is unified, incident response has to be too. A kill-switch is a design feature; what happens after it's used is a separate discipline. Firms need a single incident framework - regardless of whether the failure originated with a person, a copilot, or an agent - covering identification, severity assessment and escalation; containment mechanisms (throttling, override, temporary suspension); customer remediation playbooks; and a root-cause process that feeds findings back into Steps 1–5.

WHAT THIS MEANS FOR YOU

Every seat at the table changes

Operating model changes are only real when they change someone's job. Here is what the shift from periodic compliance to operational assurance means for each seat at the table.

Chief Risk Officer

From annual assurance to continuous assurance

The CRO's question changes from "were our AI systems compliant at the last review?" to "are they compliant right now — and how would I know within the hour if they weren't?" Risk appetite for AI becomes something monitored against, not something stated in a document.

Chief Compliance Officer

From file sampling to outcome evidence

Sampling was a proxy for evidence when full coverage was impossible. It no longer is. The CCO's mandate shifts from checking a sample of files to evidencing outcomes across every interaction and being able to show the regulator that evidence on demand.

Chief Data & AI Officer

From model deployment to lifecycle governance

Getting a model into production moves from being the finish line and becomes the starting line. The CDAO owns the system's behaviour for its entire operational life: drift, degradation, misuse, and the evidence trail that proves it remains safe.

First line

From manual QA to AI-supported assurance

Team leaders and QA functions stop reviewing a handful of interactions a week and start supervising an assurance system that reviews all of them, directing their human judgement to the cases that genuinely need it.

The Board

From quarterly reporting to operational visibility

Board packs summarising last quarter's AI risk position give way to live visibility of it. Under SM&CR, this is not a nice-to-have: accountable individuals need to see what they are accountable for, at the speed it actually happens.

The common thread: in every role, the deliverable changes from **documents about governance**
to **evidence of governance**.

HOW CAN AVENI HELP

The six steps, mapped to capability

Mapped against the six steps in this paper - where each capability tests, checks, and evidences agentic activity.

PROGRAMME (STEPS 1–6)	AVENI CAPABILITY	CONTEXT
Step 1: Establish agent approval framework	APPROVE	Tests every agent against your policies and risk tiers before it reaches a customer — turning the approval model into something enforced, not just documented.
Step 2: Encode requirements in policy	APPROVE	Converts your policy and risk appetite into rules an agent can actually be tested against, before it ever reaches a customer.
Step 3&4: Build continuous assurance, select models fit for live assessment	ASSURE FINLLM	Checks agents against those same rules continuously once they're live — spotting drift the moment it happens, not at the next scheduled review.
Step 5: Build the evidence and reporting backbone	DETECT	Assist captures and structures every interaction at source; Detect assesses outcomes against it — together, the traceable evidence base the Board, Audit, and the regulator can draw from directly.
Step 6: Unify oversight across humans and AI	APPROVE ASSURE DETECT	One consistent standard applied across every channel and interaction type — human, copilot, or agent — closing the blind spots a channel-by-channel model creates.

WHY AVENI

Specialist AI built for financial services, proven at scale

Seven years deployed | Millions of regulated interactions | 250+ clients

Aveni is building the AI value chain for financial services, combining proprietary data, domain expertise, specialist models, copilots, and continuous assurance into a single AI platform.

For seven years, Aveni has deployed AI across UK banks, wealth managers, advice firms, and insurers. Today, it supports 250+ clients through a team combining advanced AI capability with deep financial services expertise.

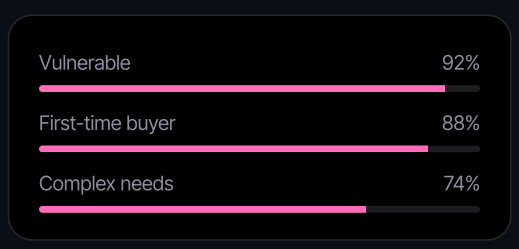
compliance

Continuous AI assurance across every customer interaction.

PRE | DURING | AFTER

Approve

TEST AI BEFORE DEPLOYMENT



Validates agent behaviour and policy alignment before deployment. Tests edge cases and compliance at scale before a real customer sees it.

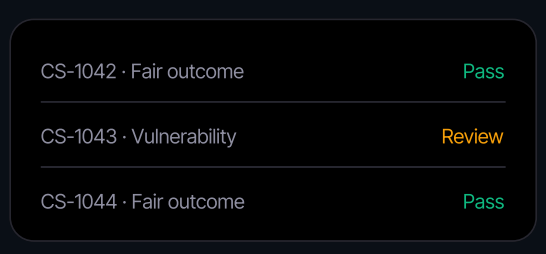
Assure

MONITOR AI CONTINUOUSLY

Live, continuous monitoring across every channel. Fast classifiers catch conduct issues and vulnerability signals in the moment.

Detect

TURN CLOSED CASES INTO EVIDENCE



Every closed case is assessed — conversations and documents together — into audit-ready evidence of fair outcomes.

Aveni Intelligence

THE INTELLIGENCE ENGINE

FINLLM • 150+ MODELS

Runs detection, reasoning, and decisioning across every interaction in real time.

TIER 1

- **Fast. Live. Continuous**

150+ models aligned to your Risk taxonomy designed to monitor agent interactions at every turn

TIER 2

- **Deep. Specialist. Context Aware SLM's**

FinLLM models 1, 8, 14bn parameters trained on extensive industry data

TIER 3

- **Generalist Hyperscaler LLM's**

Open AI, Google, Anthropic for generalist outcomes

aveni

From readiness to confidence

Aveni helps financial services firms move from Mills Review requirements to governed, evidenced, auditable agentic deployment. If the six steps in this paper resonate, we'd like to show you how we approach them in practice.

partnerships@aveni.ai