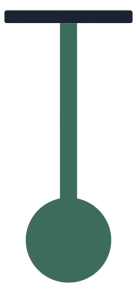
Trustethica

TRUSTETHICA | EXECUTIVE REPORT

Purpose-anchored Contextual AI Governance

Extending SAFR for higher risk AI deployments in FSI



July 2026

Purpose-anchored Contextual AI Governance – Extending SAFR for higher risk AI deployments in FSI

EXECUTIVE SUMMARY

In July 2026, the Monetary Authority of Singapore published Safeguards for Agentic Finance at Runtime (SAFR) through BuildFin.ai[1] as an industry reference approach for governing agentic financial actions before execution. SAFR places a governance checkpoint between an agent's decision and the external system on which that decision would take effect.

This checkpoint addresses a timing defect in conventional AI governance. Model validation and use-case approval occur before deployment while audit and incident review occur after the event. Neither provides assurance at the moment a machine-generated decision becomes a payment, order, filing, credit outcome, customer communication or other consequential business act.

SAFR closes that gap by requiring the proposed action to be declared in a Governance Envelope, linked to a verified Agent Identity, tested against an institutional Controls Repository, resolved through a Disposition Engine and recorded in a tamper-evident Audit Log. However, the framework draws a clear boundary.

Model guardrails shape what an agent may generate, but they do not establish that a proposed action is authorised. Settlement rails and compliance systems govern downstream execution, but they do not decide whether the agent should act in the first place. SAFR therefore sits in the decision path, after model-level controls and before execution.

Trustethica's position starts from SAFR's architecture and takes the enterprise question one level further. SAFR recognises context, mandates and institution-specific rules. The remaining implementation problem is how the institution represents, owns and continuously maintains the business context against which those rules operate.

The risk of an AI system is not determined by the agent or proposed action alone. It changes with the approved purpose, principal, caller population, data, model, tool, business process, regulatory policy requirements, downstream consequence, risk appetite and evidence obligation.

A proposed action may be technically valid and within a narrow mandate yet still be outside the authorised use of the system. Equally, a sequence of individually permissible actions may collectively move a deployment into a different business purpose or regulatory risk profile. Trustethica defines this as contextualized purpose drift, a governance failure that can occur without model drift, access failure or an obvious policy breach at the level of a single event.[2]

Trustethica addresses this by treating the authorised use as the persistent control object. Its Authorised Use Envelope is an institution-owned baseline for the business purpose and conditions under which an AI use case is approved. At runtime, live activity is evaluated against the current envelope and recorded in a Contextual Use Record that preserves the observed event, the control decision, the enterprise risk translation and the evidence generated.

[1] Monetary Authority of Singapore, Safeguards for Agentic Finance at Runtime, Version 1.0, July 2026.
[2] Purpose-anchored Contextual AI Governance, Trustethica, May 2026

Three claims that should not be collapsed

The architecture separates three claims that should not be collapsed:

- what the agent or runtime declares;
- what the enterprise has authorised; and
- what the governance system independently observed and decided.

This separation matters because the agent's declaration is evidence, not authority. SAFR itself identifies the risk that an agent-generated envelope can be internally coherent yet fail to represent the originating instruction faithfully. The authority baseline must therefore remain external to the agent, institution-owned and version-controlled.

SAFR dispositions answer the operational runtime question of execute, observe, escalate or deny. Boards, CROs and control functions need the resulting patterns translated into their existing risk taxonomy, including concentrations, repeated exceptions, model or source changes, unresolved ownership and movement against risk appetite. Runtime governance becomes credible when it combines pre-execution decisioning, enforcement at the customer boundary, accountable human escalation, longitudinal use-case control and evidence by construction.

1	SAFR changes the timing of control. The decisive governance event moves from approval and retrospective review to the point immediately before an agentic action takes effect.
2	Context is a control input. Purpose, principal, data, model, process, regulatory policy requirements, and consequence determine the enterprise meaning of an otherwise identical action.
3	Declaration and authority must remain separate. The agent can submit a structured account of its intended action, but the enterprise must own the mandate and use-case baseline against which it is assessed.
4	Approval must become a living baseline. A point-in-time approval is not evidence that live behaviour remains within scope after changes to users, data, models, tools or business process.
5	Action decisions need longitudinal interpretation. Individually permissible events can still form a pattern of contextualized purpose drift, control deterioration or risk concentration across a portfolio.
6	Observation is not the same as control. A binding control claim requires a decision that the customer enforcement boundary is obliged to honour, together with an evidentiary record of the result.

Why SAFR matters

SAFR's most important contribution is the placement of the checkpoint. Most enterprise AI controls operate in one of three places:

- Pre-deployment controls decide whether a model or use case may go live.
- Model-level guardrails constrain prompts, outputs and tool use.
- Post-event controls sample logs, investigate incidents and assess control performance.

Agentic systems expose the ungoverned space between them, the point at which a model-derived decision becomes an external act.

A model may be validated, the agent may be registered and the output may be safe in content terms, while the proposed action still exceeds delegated authority, violates a product rule, falls outside the permitted instrument set or carries a level of consequence that requires human decision. Conversely, an action may be acceptable for autonomous execution when it is reversible, low-value, in-pattern and well supported by evidence.

SAFR makes those distinctions operational. A proposed action is packaged with its action trace and relevant context, bound to a registered agent, evaluated against institution-specific controls and resolved to one of four outcomes: Auto-Execute, Observe, Escalate or Deny. The decision and its basis are then recorded. In multi-step workflows, the process repeats for each action where a prior outcome does not create continuing authority as conditions change.

> **CONTROL PRINCIPLE**
>
> A runtime log created after execution is still retrospective. SAFR's architectural advance is a decision before external effect.

This is also why runtime monitoring and runtime governance should not be treated as synonyms. Monitoring can identify a problem at machine speed and still arrive too late to prevent it. Governance requires a decision right in the execution path and an enforcement point that is required to act on the result.

The enterprise question SAFR exposes

SAFR does not ignore context. Its Governance Envelope includes action details, the action trace and context metadata while its Controls Repository can incorporate organizational policy, regulatory requirements, product rules and user-provided mandates. The enterprise gap is that institutions must make their approved context explicit, persistent and current enough to operate as a control baseline.

Agent identity answers which technical actor submitted the proposal. A mandate defines the authority delegated for a class of action. Neither, by itself, establishes the full basis on which the organisation approved the AI use. That basis may include the business purpose, accountable owner, approved users, data and source restrictions, model and tool binding, workflow location, regulatory policy requirements, downstream decision consequence, risk appetite and evidentiary obligations.

> **CORE DISTINCTION**
>
> A valid action can still be invalid in context.

Context determines the enterprise meaning

Consider a customer-service agent authorised to retrieve account information. The same retrieval action may be appropriate when answering a service query and impermissible when used to infer financial suitability or produce a credit recommendation. The identity, API call and accessed record may be unchanged. The business purpose and legal consequence are not.

The same issue arises when an internal research assistant moves into a decision workflow. Summarizing public sources may sit within an approved research purpose. Using the same output as a decisive input to underwriting, sanctions disposition or regulatory reporting changes the control obligation even if no single request appears anomalous. Context determines the enterprise meaning of the event.

The compound nature of the use also creates a longitudinal problem. A series of individual actions may each satisfy an immediate rule while the deployment as a whole shift into a new user cohort, data domain or business purpose. An action gate can resolve the local event but enterprise governance must also determine whether the pattern remains consistent with the use that was approved.

Authorised use as the control object

Trustethica's central design choice is to make the authorised use, not the model, agent or inventory record, the persistent control object.

The Authorised Use Envelope is a governance contract in the operational sense. It records the conditions under which the institution has approved a specific use of AI and provides the baseline against which runtime activity is assessed. It is institution-owned, versioned and connected to accountable decision-makers.

An effective envelope should cover the following control domains:

- authorised and prohibited business purposes, including the relevant business process and downstream consequence;
- accountable business, risk and technical owners, together with approved users, principals, agents and service identities;
- approved models, tools, integration paths, data classes, sources and jurisdictions;
- enterprise risk mapping, appetite thresholds, materiality factors and control outcomes;
- human review, escalation, suspension, attestation and change-management obligations; and
- the evidence that must be preserved to demonstrate which authority baseline and rule set applied at the time of each event.

The envelope becomes a living control target when runtime events are bound to the version in force and material changes trigger review. A model substitution, new data source, expanded user cohort, changed workflow, new regulatory policy requirement or revised risk threshold can alter the basis on which the use was approved. The governance record should change before the operational reality outlives the approval.

This distinction changes the governance objective. The institution is no longer asking only whether a model or agent is behaving as designed. It is asking whether the current use, in its present context, remains inside the authority and risk boundary that the institution has accepted.

Authority must remain external to the agent

SAFR explicitly recognises that an agent-generated envelope can be internally coherent while failing to represent the originating instruction faithfully. The declaration therefore requires authentication against its origin. The broader control consequence is clear; the actor cannot be the sole source of its own authority.

An institution-owned envelope also allows the enterprise to revise a mandate, risk threshold or data boundary without relying on the agent to restate the change correctly. Binding the runtime record to the applicable envelope version allows audit and risk functions to reconstruct not only what happened, but why the organisation considered the event permissible, observable, escalated or prohibited at that time.

Human oversight that is operational, not ceremonial

SAFR treats human escalation as an operational design problem, not a statement of principle. Three dimensions determine whether review is substantive:

- the volume that reviewers can realistically process;
- the time within which a decision must be made; and,
- the authority of the reviewer to approve, modify or reject the action.

That framing is important because "human in the loop" is often used without defining the loop. A notification does not create oversight if it arrives after execution, lacks the evidence needed for decision, has no response deadline, or is routed to a person who cannot change the outcome. An unmanageable escalation queue is not conservative governance; it is an undisclosed fail-open posture.

Trustethica extends the action-level escalation model into use-case accountability. Runtime advisories and exceptions can be tied to named owners, envelope versions, source events and required attestations. Lapse consequences, escalation paths and decision rights become part of the control design rather than administrative follow-up.

The objective is not to require a human to review every event. Machine-speed governance depends on humans defining the authority boundaries, thresholds and exception classes in advance, then intervening where ambiguity, materiality or control deterioration warrants judgement. The runtime system scales the policy but accountable humans retain the decision rights.

The operating model should also preserve lines of defense. First-line owners remain accountable for the AI use and its controls. Second-line functions set or challenge risk parameters and escalation standards. Internal audit provides independent assurance. A runtime governance platform should evidence those responsibilities, not blur them.

Executive implementation agenda

The most effective implementation path is not an enterprise-wide inventory exercise. It is a bounded deployment against one consequential use case where authority, control decisions and evidence can be tested in the operating environment.

PHASE	EXECUTIVE DECISIONS	EVIDENCE OF PROGRESS
Days 0-30: Define	Select one high-consequence use; name business, risk and technical owners; define authorised and prohibited purposes; identify the pre-execution or pre-release control point.	Approved Authorised Use Envelope, mapped decision rights, initial control set and agreed success criteria.
Days 31-60: Instrument	Bind live events to the use case; configure identity, data, model and purpose checks; calibrate dispositions and human review; define enforcement posture.	Resolved runtime events, tested control outcomes, reviewer workflow, failure-mode results and initial evidence chain.
Days 61-90: Operate	Run against real or near-real activity; assess drift, review capacity and risk translation; remediate control gaps; decide whether and how to scale.	Evidence pack, risk-appetite view, accountable exception decisions, control limitations and scale recommendation.

Questions for the executive sponsor

01 What exactly is the authorised use of a particular AI deployment?

02 Which event must be assessed before an external effect, consequential output or regulated workflow step occurs?

03 Which enterprise-owned record contains the current purpose, scope, owner, data, model, risk and evidence baseline?

04 What happens when identity, use-case resolution, model version or evidence is ambiguous?

05 Which outcomes are advisory and which are binding at the enforcement boundary?

06 Can human reviewers process the expected volume within the required time and exercise real decision authority?

07 Can each board-level risk signal be traced to the underlying event, rule, envelope version and owner decision?

08 What change triggers invalidate the prior approval and require re-attestation or suspension?

Success should be defined in operational terms where every governed event resolves to a current authority baseline or an explicit unresolved outcome, control decisions are enforced as configured, material divergence reaches an accountable owner, and the organisation can produce a coherent evidence record without reconstructing it after the fact.

Conclusion

SAFR provides a credible and timely foundation for agentic AI governance. It identifies the correct intervention point, separates model guardrails from action authority, makes human escalation operational and requires a consistent record of the decision. It is intentionally a starting point that institutions must implement through their own controls and governance arrangements.

Trustethica's differentiated contribution is to operationalize the part SAFR leaves to the institution, a standing, enterprise-owned definition of the authorised AI use and a continuous comparison between that authority baseline and live behaviour.

The SAFR Governance Envelope carries the proposed action and its supporting trace. The Trustethica Authorised Use Envelope carries the institution's current authority, risk and evidence baseline. The Contextual Use Record preserves what was observed, decided, enforced and translated into enterprise risk. Together, these artefacts connect machine-speed action control to the level at which enterprise accountability actually sits.