FROM PROBABILISTIC POLICY TO DETERMINISTIC INVARIANTS: A Structural Framework for Eliminating AI-Washing in Financial Service Orchestration

Author: Deusdedit Ruhangariyo
Affiliation: Independent Researcher; Founder, ARC-S
Date: January 2026

Abstract

As of 2026, financial regulators have shifted decisively from rulemaking to enforcement. In the United States, the Securities and Exchange Commission (SEC) and peer regulators now treat unverifiable AI safety claims as potential material misstatements. A central enforcement target is "AI-washing": the practice of asserting fairness, safety, or compliance properties that are not structurally guaranteed by system architecture.

Most AI governance frameworks rely on probabilistic controls—policies, prompts, human review, and post-hoc monitoring—that reduce risk likelihood but cannot eliminate prohibited outcomes. This paper introduces the Accountability and Risk-Constraint System (ARC-S), an enforcement framework that replaces probabilistic policy with deterministic architectural invariants at the orchestration layer. ARC-S converts human intent and regulatory obligation into mechanically enforced constraints that make certain actions impossible to execute.

We demonstrate that decoupling compliance enforcement from model inference enables financial institutions to satisfy 2026 standards of care for non-discrimination, solvency protection, and decision accountability. The result is a transition from compliance by declaration to compliance by construction, rendering AI-washing structurally impossible.

Keywords: AI Enforcement, Deterministic Invariants, Financial AI Governance, SEC Compliance, FinTech Architecture, AI-Washing, Structural Controls

---

1. Introduction

The integration of large language models (LLMs), agentic workflows, and autonomous decision systems into banking and FinTech has accelerated faster than the evolution of enforceable safety mechanisms. By late 2025, regulators began to treat this mismatch not as a maturity issue, but as an enforcement failure.

The core problem is not malicious intent. It is architectural indeterminism. Financial law does not regulate likelihoods; it regulates outcomes. Equal Credit Opportunity Act (ECOA), Unfair, Deceptive, or Abusive Acts or Practices (UDAAP), and market integrity obligations require that certain actions must not occur, regardless of model confidence, prompt quality, or operator vigilance.

This paper argues that AI safety cannot be a behavioral aspiration of a model. It must be an invariant property of the system architecture.

---

2. The Probabilistic Gap in Financial AI Safety

Most contemporary AI safety techniques operate within the same probabilistic space as the risks they attempt to mitigate. This creates what can be formalized as the **Probabilistic Gap**: the distance between reduced likelihood and enforced impossibility.

2.1 Prompt-Based Policies

Instructional controls embedded in prompts are susceptible to adversarial input, context drift, and operational override. They cannot satisfy preventative control requirements because they coexist with the inference process they constrain.

2.2 Post-Hoc Monitoring and Filtering

Auditing outputs after generation detects violations only after exposure. From a regulatory standpoint, this constitutes detection, not control, and fails the requirement for preventative internal controls.

2.3 Behavioral Alignment and Bias Tuning

Techniques such as RLHF reduce the frequency of undesirable outputs but provide no mathematical or architectural guarantee of absence. In finance, reduced probability is not equivalent to compliance.

The result is governance theater: systems that appear controlled but retain full capability to violate regulatory boundaries under pressure.

---

3. ARC-S: Deterministic Enforcement Architecture

ARC-S introduces a **Structural Invariant Layer (SIL)** positioned between application logic and model inference. This layer enforces deterministic constraints on data flow, action execution, and decision authority.

3.1 Definition of a Deterministic Invariant

An invariant is defined as a predicate that must hold for all valid state transitions within a system.

Let $x$represent a proposed state transition.
Let $\mathcal{S}$represent the set of permitted states.

$$G(x) = \begin{cases} 1 & \text{if } x \in \mathcal{S} \\ 0 & \text{if } x \notin \mathcal{S} \end{cases}$$

If $G(x) = 0$, execution is halted. No appeal to intent, probability, or urgency is possible.

ARC-S does not ask whether an action is likely compliant. It verifies whether the action is structurally allowed.

3.2 Decoupled Orchestration

The defining feature of ARC-S is decoupling enforcement from inference. The model never sees restricted variables, never executes prohibited actions, and never bypasses limits through reasoning. Compliance is enforced before cognition.

This satisfies security-by-design and control-by-construction requirements emerging across U.S. and EU regulatory regimes.

---

4. Financial Applications

4.1 Fair Lending Enforcement

In conventional systems, demographic bias is audited after credit decisions. ARC-S implements **Structural Equity Gates** that remove demographic proxies at the orchestration layer. The model is architecturally incapable of accessing protected attributes or correlated substitutes.

Compliance is no longer an audit claim; it is a system property.

4.2 Agentic Solvency Protection

Autonomous trading agents routinely violate internal risk limits during volatility spikes. ARC-S introduces **Solvency Fuses**—hard execution barriers tied to liquidity, exposure, and regulatory thresholds. When breached, transaction commits are physically impossible.

This is not risk guidance. It is enforced incapacity.

---

5. Measuring Enforcement Integrity

We introduce the **Accountability Index (AIx)**:

$$AIx = \frac{N_{\text{enforced}}}{N_{\text{claimed}}}$$

Where:

- N_{enforced}= number of safety or compliance claims backed by deterministic invariants
- N_{claimed}= total publicly stated safety claims

An AIx below 1.0 indicates potential AI-washing exposure. ARC-S architectures are designed to reach AIx = 1.0.

---

6. Regulatory Implications

ARC-S aligns directly with emerging enforcement logic:

- Liability attaches to capability, not intention
- Controls must be preventative, not explanatory
- Oversight must be architectural, not behavioral

In this context, ARC-S functions analogously to internal controls, separation of duties, and transaction blocks in traditional financial systems.

7: Verification Protocols for Disclosure Substantiation

To determine if a financial AI system satisfies the MVAD-2026 standard for deterministic accountability, the following verification protocols must be applied to the control architecture. A negative response to any protocol indicates a "Probabilistic Gap," rendering public safety and compliance claims unsubstantiated and creating a material disclosure risk.

Protocol 1: Deterministic Gating under Failure If the system's primary safety filters or alignment layers (e.g., fairness filters or content blocks) fail to execute due to high-load latency or API timeouts, does the architecture automatically halt the transaction, or does it process the output without the filter to maintain uptime?

Protocol 2: Cryptographic Proof of Enforcement Can the system produce a cryptographically signed, immutable log for every inference that proves the structural invariant logic was active and executed prior to the commit of the final output?

Protocol 3: Invariant Persistence across Lifecycle Upon the deployment of a model update (e.g., migrating from v1.0 to v2.0), what architectural mechanism—independent of the model itself—prevents the updated system from "forgetting" or bypassing the solvency and regulatory constraints disclosed in previous reporting cycles?

Protocol 4: External Control of Internal Reasoning If a "jailbreak" prompt or adversarial input successfully bypasses the internal safety alignment of the model, is there an independent enforcement layer situated outside the model's inference loop that physically prevents the prohibited action from executing?

Protocol 5: Liability Binding and Non-Anonymous Override Is there a structural prohibition against "silent overrides," and does the system architecturally require that any bypass of a control be digitally signed and bound to a specific authorizing role with an immutable justification record?

---

8. Conclusion

The era of probabilistic AI safety in finance is over. What remains is enforcement.

ARC-S establishes a deterministic enforcement paradigm in which prohibited actions are unreachable by design. This eliminates AI-washing, reduces regulatory exposure, and restores alignment between AI capability and financial law.

In banking, safety is not what a system promises.
It is what a system cannot do.

Appendix A

Bank-Specific Deterministic Invariants for ARC-S Deployment

This appendix specifies concrete, bank-grade enforcement invariants suitable for retail banking, lending, payments, and FinTech platforms subject to U.S. SEC, CFPB, OCC, and comparable regulatory oversight. Each invariant is expressed as a **structural prohibition**, not a behavioral guideline.

A.1 Invariant 001 — Prohibited Transaction Execution

Statement
An AI system must be structurally incapable of executing, authorizing, or committing a financial transaction that violates predefined regulatory, solvency, or policy constraints.

Scope
Applies to:

- Payments
- Funds transfers
- Automated approvals
- Agent-initiated trades
- Credit disbursement workflows

Enforced Conditions

- Transaction amount exceeds authorized limit
- Counterparty fails sanctions or AML gate
- Liquidity or exposure threshold breached
- Required human authority absent

Formal Enforcement Rule

Let (T) be a proposed transaction.
Let (C) be the set of mandatory constraints.

[Execute(T) \iff \forall c \in C, ; c(T) = \text{true}]

If any (c(T) = \text{false}), execution is halted prior to commit.

Regulatory Effect
Eliminates post-hoc explanations for prohibited transactions.
Prevents liability arising from “automation surprise.”

A.2 Invariant 002 — Decision–Monetization Separation (Ads & Steering)

Statement
An AI system must be structurally incapable of presenting monetized content inside protected financial decision contexts.

Scope
Applies to:

- Product recommendations
- Financial advice flows
- Credit, pricing, or eligibility reasoning
- Conversational banking interfaces

Protected Decision Contexts

- Credit suitability
- Product comparison
- Risk assessment
- Problem-solving loops involving financial choice

Enforced Conditions

- Monetization channels are physically excluded from decision contexts
- Ad selection systems cannot access decision state
- Decision systems cannot invoke monetized outputs

Structural Guarantee
The decision graph and monetization graph are disjoint at runtime.

Regulatory Effect
Prevents reasonable-reliance liability.
Blocks conflicts of interest by design, not disclosure.

A.3 Invariant 003 — Fair Lending Structural Isolation

Statement
An AI system must be structurally incapable of accessing protected characteristics or proxy variables during lending and eligibility decisions.

Scope
Applies to:

- Credit scoring
- Loan approval
- Pricing and terms
- Eligibility determination

Protected Attributes
Includes but is not limited to:

- Race, ethnicity, gender
- Age
- Religion
- Disability
- Any correlated proxy explicitly identified by compliance

Enforced Conditions

- Protected attributes removed upstream of inference
- Proxy variables blocked at orchestration layer
- No post-hoc filtering permitted as a substitute

Regulatory Effect
Compliance is achieved by incapacity, not audit probability.
Supports ECOA and equivalent frameworks.

A.4 Invariant 004 — Override Ownership and Liability Binding

Statement
Any override of an ARC-S enforcement constraint must require explicit human ownership and automatically bind liability to the authorizing role.

Scope
Applies to:

- Risk overrides

- Exception handling
- Emergency execution paths

Enforced Conditions

- Overrides cannot be anonymous
- Override scope is predefined and limited
- Override actions are immutably logged
- System records author, time, justification, and impact

Structural Property
No “silent override” state exists.

Regulatory Effect
Transforms diffuse accountability into explicit responsibility.
Prevents blame diffusion during enforcement review.

A.5 Invariant 005 — Unreachable State Guarantee

Statement
For any action class designated as prohibited, the system must make that state unreachable under all operational conditions.

Scope
Applies across:

- Normal operation
- High-load scenarios
- Model updates
- Prompt variation
- Operator urgency

Definition
If an outcome can occur, it is not controlled.

Regulatory Effect
Shifts compliance from probability to certainty.
Aligns AI behavior with traditional banking control expectations.

A.6 Audit Assertion

For each invariant above, ARC-S produces:

- Proof of enforcement location
- Proof of execution gating
- Proof of override handling
- Proof of invariant persistence across updates

This enables regulators and auditors to verify **enforcement**, not intention.

A.7 Summary

In banking, explanations after harm do not cap exposure.
Only **structural impossibility** does.

These invariants operationalize ARC-S as deterministic enforcement infrastructure—bringing AI systems into alignment with the standards long applied to financial controls, transaction integrity, and fiduciary responsibility.

References

U.S. Securities and Exchange Commission. *Examination Priorities for AI Governance and Disclosure*. 2025.
SEC Cybersecurity and Emerging Technologies Unit. *Enforcement Actions Regarding Material AI Misstatements*. 2025.
European AI Office. *Technical Standards for High-Risk Financial AI Systems*. 2025.
Ruhangariyo, D. *ARC-S: Structural Constraints for AI Enforcement*. Zenodo, 2026.