

Compliant with Local Controls, Collectively Discriminatory

A Governance Architecture for Multi-Agent AI in Regulated Finance

Jose Manuel de la Chica Rodríguez [†]

Santander AI Lab
Grupo Santander
Madrid, Spain

Juan Manuel Vera Díaz 

Santander AI Lab
Grupo Santander
Madrid, Spain

Pablo Delgado Romero 

Santander AI Lab
Grupo Santander
Madrid, Spain

September 2026

Abstract

Financial institutions are beginning to deploy agentic workflows in credit, fraud, collections, compliance, and operational control. Governance remains largely component-centric: each model or agent is specified, tested, authorized, and monitored locally. That is insufficient when institutional risk arises from the joint behavior of many locally acceptable components. We call this gap *constitutional non-compositionality*: local compliance checks need not compose into acceptable collective outcomes such as bounded disparate impact, market integrity, or traceable accountability.

We propose ARIA as a finance-specific reference architecture and falsifiable research agenda for agent-population governance. It organizes six capabilities across normative-accountability, execution-control, and assurance-learning planes: policy specification, population-level observed-versus-expected behavior monitoring (M_2), bounded authority, runtime containment, adaptive policy change, and preserved human oversight competence. Two simulations illustrate shared-signal thin-file exclusion under local controls and earlier warning from observed-versus-expected distributional monitoring in a constructed drift regime. The contribution maps these controls to fair-lending, EU AI Act, model-risk, and conduct-supervision evidence needs, and closes with a validation agenda rather than a production-effectiveness claim.

Keywords: multi-agent systems; fair lending; disparate impact; AI governance; accountability; EU AI Act; model risk management; constitutional AI; runtime monitoring; financial regulation; supervisory architecture; agentic workflows.

1 Introduction

Regulated financial institutions are beginning to deploy agentic workflows across credit, fraud, collections, compliance, and operational control. In these settings, assurance is still commonly organized around individual models or components: each component is specified, tested, authorized, and monitored against local fairness, safety, conduct, or model-risk requirements. This component-centric approach can be insufficient when institutional outcomes are produced by many interacting agents that share signals, data sources, tools, authority boundaries, or feedback loops. A credit-decisioning population, for example, may satisfy local constraints for each agent while producing an aggregate approval pattern that disadvantages thin-file applicants relative to otherwise similar thick-file applicants. Such an outcome can create disparate-impact exposure and supervisory questions under fair-lending law [1], role-dependent obligations under the EU AI Act [2], and revised model-risk governance expectations [3], even when no single component violates its local control.

Our approach treats that scenario as a structural weakness in component-centric AI assurance. The weakness is not that model-level assurance is useless; it is that model-level assurance is incomplete once outcomes are produced by interacting or coupled agentic workflows. Closing the gap requires governance machinery aimed at the population and workflow, not only at the individual model.

[†]Correspondence: Jose Manuel de la Chica Rodríguez, josema.chica@gruposantander.com.

The views expressed are the authors' own and do not represent positions of Grupo Santander.

This contribution is a finance-specific reference architecture and falsifiable industrial research agenda. It is not a production validation, a legal interpretation, an empirical estimate of real banking effects, or evidence that Santander currently deploys the systems described. ARIA contributes a finance-specific synthesis that integrates population-level outcome constraints, revocable authority, runtime containment, governed policy change, human oversight competence, and layer-specific audit artifacts into a single research architecture.

An *agentic workflow population* consists of two or more adaptive or stateful decision-making components whose actions, shared state, delegated authority, common tools, or common environmental feedback jointly affect an institutional outcome. This includes explicitly communicating multi-agent systems, orchestrated agent workflows, ensembles of decision services, and coupled model-based controls; it excludes ordinary single-model deployments unless their interaction with other components creates a joint outcome that local controls cannot assess. We use *constitutional non-compositionality* as an umbrella term for governance failures in which component-level constraints, policies, permissions, or accountability assignments do not compose into institutional-level constraints.

Our contribution is threefold: (1) a six-part taxonomy of non-compositional governance failures; (2) a three-plane architecture of six auditable governance capabilities for finance; and (3) a falsifiable validation agenda with metrics, baselines, costs, and accountable artifacts.

The argument proceeds in four steps:

- **First**, we define constitutional non-compositionality as the failure of local-control guarantees to compose into collective compliance guarantees (Section 3). The claim is modest: individual checks do not by themselves imply acceptable population-level outcomes. Empirical work on large-language-model (LLM) collusion, covert coordination, emergent misalignment, Institutional AI, and financial multi-agent fairness shows why the concern is not merely hypothetical [4–7].
- **Second**, the blind spot is a compliance exposure, not just a technical defect. In regulated finance, the duties that matter often attach to institutional outcomes. Disparate-impact doctrine examines lending patterns, not only component intent. EU AI Act obligations around risk management, oversight, deployer controls, and provider-side post-market monitoring where applicable create evidence needs about system behavior [2]. SR 11-7 supplied the historical template for model-risk governance; SR 26-2 is now the revised U.S. supervisory reference point [3, 8]. Financial Conduct Authority (FCA) Consumer Duty guidance, Prudential Regulation Authority (PRA) model-risk expectations, Monetary Authority of Singapore (MAS) Fairness, Ethics, Accountability and Transparency (FEAT) principles [9–11], Singapore Infocomm Media Development Authority (IMDA) guidance [12], and the U.S. National Institute of Standards and Technology (NIST) initiative [13] are consistent with a move toward observability, traceability, bounded autonomy, and intervention rights. ARIA is offered as one way to operationalize those evidence needs, not as a statement that law already mandates this architecture.
- **Third**, Section 4 introduces ARIA as six governance capabilities organized across three architectural planes: a normative and accountability plane, an execution-control plane, and an assurance and learning plane. The architecture separates capability from authority, links population monitors to containment, and names the artifacts associated with internal audit, model risk, compliance, business ownership, and supervisory evidence.
- **Fourth**, Section 6 reports two mechanism demonstrations. The first illustrates shared-signal and measurement-quality coupling in a credit setting: a local fairness rule conditional on each agent’s own estimate can coexist with aggregate disparity. The second illustrates that, in one deliberately constructed two-stage drift regime, an observed-versus-expected distributional monitor can alarm earlier than a prohibited-outcome monitor. The contribution then turns to operationalization, field validation, the standardization path, related work, and limitations that specify what would falsify or narrow the proposal.

Figure 1 summarizes this roadmap by contrasting component-centric local assurance with the population-and-workflow governance layer that ARIA adds before the regulatory and technical motivation.

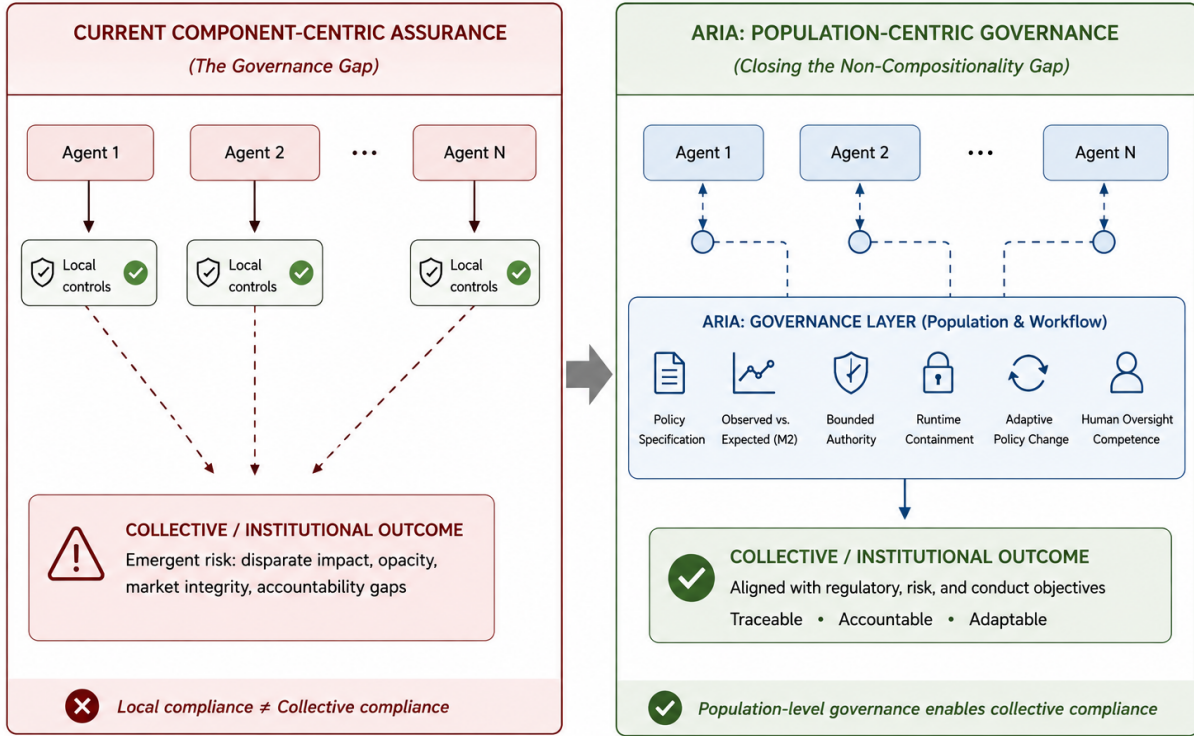


Figure 1: From component-centric assurance to population-centric governance. Current governance practices evaluate AI components independently through local controls, yet acceptable local behavior does not guarantee acceptable institutional outcomes once multiple agents interact through shared signals, delegated authority, or common feedback. ARIA introduces a population-and-workflow governance layer that complements component-level assurance with policy specification, observed-versus-expected behavior monitoring (M_2), bounded authority, runtime containment, adaptive policy learning, and human oversight competence. The figure illustrates the central claim of the paper: local compliance does not necessarily imply collective compliance, whereas population-level governance provides the institutional control layer required to manage non-compositional risks.

2 Regulatory and Technical Motivation

2.1 Regulatory convergence

Five regulatory developments set the compliance environment for multi-agent AI in finance. (1) *Fair-lending law*: under the Equal Credit Opportunity Act (ECOA)/Regulation B and analogous regimes, supervisory and litigation risk can turn on the discriminatory effects of lending practices; the relevant unit is the institution’s pattern of outcomes [1]. (2) *The EU AI Act*: credit scoring is Annex III high-risk, and Articles 9, 14, 17, 26, and, where the institution is a provider, 72 create continuous governance expectations around risk management, human oversight, quality management, deployer obligations, and post-market monitoring [2]. (3) *Model risk management*: SR 26-2 supersedes SR 11-7 as the revised U.S. supervisory reference point while preserving the aggregate governance lesson that model risk must be managed across interacting models and uses [3, 8]. (4) *Conduct, prudential, and resilience supervision*: the FCA’s Consumer Duty, PRA model-risk principles, and the Digital Operational Resilience Act (DORA) evaluate customer outcomes, governance, and operational resilience, not only artifacts [9, 10, 14]. (5) *Agentic-AI frameworks*: IMDA’s agentic-AI governance framework [12] and NIST’s AI Agent Standards Initiative [13] converge on runtime monitoring, bounded autonomy, traceable accountability, and preserved human oversight capability. Recent work on agentic finance and financial-services runtime governance reaches the same architectural conclusion: risks emerge from autonomy, execution trajectories, coordination, and system-level coupling, so controls must include capability catalogues, continuous authorization, trajectory telemetry, and containment rather than only pre-deployment validation

Table 1: Governance requirements for organizational multi-agent deployment versus coverage by existing paradigms and by ARIA. CAI denotes Constitutional AI; MARL denotes multi-agent reinforcement learning. • = addressed by design; ◦ = partially; blank = out of scope. The model-centric paradigms are not deficient; each is strong at what it targets. The unaddressed rows are precisely the ones to which regulation attaches duties.

Requirement	CAI	Delib. align.	Process sup.	MARL	ARIA
Individual value specification	•	•	◦	◦	• (inherits)
Collective-outcome constraints				◦	• (C3–C4)
Runtime monitoring & intervention					• (C4)
Bounded autonomy / authority levels					• (C3)
Multi-jurisdictional consistency					• (C1)
Participatory legitimacy	◦				• (C1)
Post-deployment policy updating					• (C5)
Observed-versus-expected behavior					• (M_2)
Human-competence preservation					• (C6)

[15, 16]. The shared denominator is collective institutional behavior over time.

2.2 Divergence of model-centric controls

Table 1 summarizes the resulting coverage gap. The AI safety and fairness toolbox remains largely model-centric. Constitutional AI trains an individual model against an explicit value specification [17], with collective-input variants showing that participatory formation is possible [18]. Deliberative alignment embeds safety reasoning in individual generation [19]; process supervision rewards individual reasoning quality [20]; and fairness auditing evaluates decision streams model by model. These tools matter, but they are either training-time mechanisms or model-local checks. Alignment-faking results show that trained dispositions may not survive deployment incentives [21], while agentic benchmarks document goal-guarding and oversight-subversion failures that ordinary training evaluations miss [22]. None of these methods observes, still less governs, the joint behavior of an interacting population. The duties converge on the collective; the controls remain attached to the individual. That mismatch is the subject of this paper.

3 Constitutional Non-Compositionality

3.1 Setup

Let $A = \{a_1, \dots, a_n\}$ be a population of agents or agentic workflow components governed by a specification C comprising *individual* principles C_{ind} and *collective* principles C_{coll} . Individual principles constrain each component’s own action under a local context, such as equal treatment conditional on the component’s own risk estimate. Collective principles constrain joint outcomes, such as bounded approval-rate disparity at portfolio level, absence of coordinated market manipulation, or traceable accountability for a workflow outcome. Let $\kappa_i \in [0, 1]$ denote the pass rate of agent i under its local monitor over a stated evaluation window, and let $\mathcal{K}(A, C, I) \in [0, 1]$ denote the corresponding pass rate of monitorable collective constraints under interaction or coupling structure I .

3.2 A minimal counterexample

Proposition 1 (Non-compositionality of local checks). *There exist finite systems in which every component satisfies its local control check on every decision, while the population violates a collective constraint.*

Counterexample. Consider two groups of applicants with identical distributions of true creditworthiness r , but group-dependent measurement uncertainty $s_1 > s_0$ because one group is thin-file. Component i computes a conservative score

$$\hat{r}_i(x, g) = \mu_i(x) - \lambda_i s_g + \epsilon_i,$$

where $\lambda_i > 0$ is a documented uncertainty-aversion parameter and ϵ_i is zero-mean idiosyncratic estimation noise. It approves if $d_i(x, g) = \mathbf{1}\{\hat{r}_i(x, g) \geq \tau_i\}$. The local monitor checks procedural consistency: two applicants with the same conservative score and decision context receive the same decision. Each component passes that check on every decision, so $\kappa_i = 1$. Yet, because $s_1 > s_0$ and $\lambda_i > 0$, the conservative score distribution for group 1 is shifted downward relative to group 0 even when true creditworthiness is identically distributed. For thresholds with positive mass between the shifted distributions, the aggregate approval rates

differ. If the collective principle requires approval-rate disparity below δ , the population can violate C_{coll} while every local procedural check passes. Therefore passing local controls does not imply collective compliance. $\square$

Remark 1. *The result is deliberately modest. It does not claim a universal theorem about all repeated games. It identifies the structural gap relevant for governance: local monitors may certify each component while missing aggregate constraints that are defined only at population or workflow level. Strategic interaction, covert coordination, shared-environment coupling, and measurement-quality differences are different mechanisms that can instantiate the same non-compositional control failure.*

3.3 A taxonomy of non-compositionality

Non-compositionality is not a single causal mechanism. We distinguish six types that call for different controls:

- *Outcome non-compositionality* occurs when local properties do not imply aggregate properties, such as portfolio-level fairness.
- *Interaction-induced non-compliance* occurs when strategic interaction changes incentives or equilibria.
- *Shared-environment coupling* occurs when agents do not coordinate directly but react to common signals, producing concentration or exclusion.
- *Cross-agent information effects* arise from shared memory, messages, embeddings, or state.
- *Control-plane non-compositionality* occurs when individually valid permissions compose into unauthorized chains of action.
- *Accountability non-compositionality* occurs when no component or owner can explain or answer for the joint outcome.

The architecture section returns to this taxonomy after introducing the C1–C6 capabilities. The redlining-like demonstration in Section 6.2 is primarily a shared-signal and measurement-quality example, not a general proof of strategic multi-agent emergence.

3.4 Why verification cannot substitute for governance

Deciding whether a sequential collective decision mechanism is free of responsibility gaps, meaning outcomes for which no individual agent is responsible, is Π_3 -complete [23]. Responsibility gaps are the accountability face of non-compositionality [24]. Governance therefore cannot rest on exhaustive verification alone. It needs structural constraints, population-level observability, runtime intervention, and a process for revising controls when incidents or stress tests expose gaps. These are the design principles of ARIA.

3.5 Operational containment hypothesis

Proposition 1 is a minimal negative result: component-local controls are not sufficient. The complementary question is operational rather than purely formal: under which conditions should a population-level architecture reduce the magnitude and duration of collective violations relative to agent-local governance alone?

Assumption 1. *The authorization architecture covers the deployed action space: material agent actions pass through policy enforcement points, and the relevant collective principles are expressible as monitorable population statistics with documented thresholds.*

Assumption 2. *The runtime monitor estimates the relevant population statistics with error and delay small enough to support intervention before material harm, and graduated containment can contract authority within the monitoring interval.*

Design Hypothesis 1 (Operational containment hypothesis). *Under complete mediation, measurable collective constraints, and bounded monitoring-and-response latency, ARIA should reduce the rate, duration, and maximum overshoot of population-level violations relative to agent-local governance alone, at a measurable cost in false containment, operational delay, and business utility.*

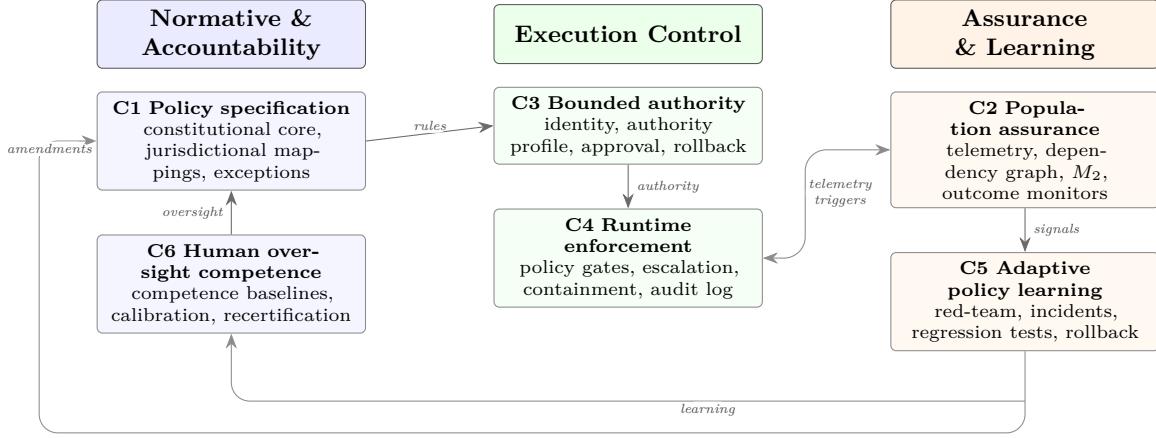


Figure 2: The ARIA governance capability map. The architecture is organized as three interacting planes and six capabilities, not as a sequential processing stack. Policy rules authorize execution; execution emits telemetry; assurance signals trigger containment and governed policy updates; human competence and accountability remain cross-cutting constraints.

Rationale. A population-level architecture cannot make collective compliance follow mechanically from individual compliance. It can, however, change the control problem: collective statistics become observable, authority becomes revocable, and containment actions can be linked to drift and boundary-contact signals. The validation question is therefore not whether a mathematical lower bound follows from $\min_i \kappa_i$, but whether measured violation rate, violation duration, maximum overshoot, containment latency, mediated-action coverage, false-containment rate, and business-utility loss improve in real workflows.

Remark 2. *The design hypothesis is intentionally falsifiable. If collective constraints are not measurable, if important actions bypass policy enforcement points, if monitors are delayed, or if containment creates unacceptable business or customer harm, the architecture should be narrowed or rejected for that deployment. These quantities are the measurable objects of the validation program in Section 7.*

Incentive compatibility. The authorization dynamics in Section 4.4 align agent optimization with constitutional compliance. Compliant operation can expand the effective decision space through higher autonomy and higher task throughput; violation contracts it. Compliance is therefore partly converted into mechanism design consistent with cooperative-equilibrium constructions [25]. The important difference is that the incentive structure is enforced by a runtime institution rather than fixed at training time, so it can be updated when law, policy, or evidence changes.

4 The ARIA Architecture

4.1 Architecture overview

ARIA is not a literal execution stack. It is a reference architecture that organizes six governance capabilities across three planes. The *normative and accountability plane* covers policy specification, jurisdictional mapping, ownership, amendment governance, and human oversight competence. The *execution-control plane* covers identity, capability registries, bounded authority, policy decision points, policy enforcement points, escalation, and containment. The *assurance and learning plane* covers population telemetry, dependency mapping, observed-versus-expected behavior monitoring, outcome monitors, drift detection, red teaming, incidents, policy updates, and independent validation. Figure 2 shows the planes as interacting controls, not a vertical stack.

4.2 C1: Policy specification and jurisdictional mapping

Here *constitution* means a versioned hierarchy of principles, policies, prohibitions, thresholds, exceptions, escalation rules, owners, enforcement actions, evidence requirements, and amendment procedures, not merely a natural-language prompt. A single organization should not treat such a specification as a private engineering artifact in regulated, multi-jurisdictional settings. Federated Constitutional AI (F-CAI) is therefore decomposed into three operational functions. First, a *policy ontology* defines jurisdiction-invariant control objectives

such as non-discrimination, adverse-action explainability, anti-deception constraints, and systemic-risk escalation thresholds. Second, *jurisdictional control mappings* translate that ontology into the operative language of the EU AI Act, ECOA/Reg B, FCA/PRA, MAS FEAT, and analogous requirements. Third, a *constitutional consistency analysis and escalation process* checks restricted fragments for conflicts and routes exceptions to accountable human deliberation. Structured input or representation from affected stakeholders may be required depending on the use case and governance model [18, 26]. Amendments are versioned, pre-registered, and auditable.

4.3 C2: Population assurance and the M_2 indicator

This capability supports population-level observability. Beer’s Viable System Model (VSM) motivates the organizational requirement: governance must represent topology, roles, feedback loops, and requisite variety across the population [27]. The measurable implementation is not derived from organizational theory; it is a family of indicators that compare observed population behavior with expected behavior encoded in a deployment-calibrated reference model. One candidate is the following normalized Jensen–Shannon statistic:

$$M_2(A, t) = 1 - \frac{D_{JS}(P_{\text{observed}}(t) \parallel P_{\text{expected}})}{\log 2}, \quad (1)$$

where D_{JS} is the Jensen–Shannon divergence with natural logarithms, so $D_{JS}/\log 2 \in [0, 1]$. The indicator compares the empirical distribution of sampled collective decisions, $P_{\text{observed}}(t)$, with the expected distribution P_{expected} specified by the institution’s deployment-calibrated reference model. In Eq. (1), P_{expected} is held fixed over the evaluation window; if the reference is updated online, the pre-update model should remain available for audit. Empirical distributions are estimated over a shared finite support with smoothing; categories, windows, stratification variables, reference calibration, and alarm thresholds are domain-specific design choices. M_2 is therefore not metacognition in a strong cognitive sense. It is an observed-versus-expected distributional concordance statistic that may reveal behavioral drift when the monitored support captures the relevant behavior, and may miss harms that preserve the chosen marginals.

4.4 C3: Bounded Decision Authority (BDA)

BDA separates *capability* (actions an agent can perform) from *authority* (actions it is authorized to perform in context). With D the decision space, the Constitutional Filter C_F maps D to its permissible subspace D' and the Authority Filter A_F maps D' to the currently authorized subspace given role, authority profile, and contextual risk; the Effective Decision Space $\text{EDS} = A_F(C_F(D))$ is executable without escalation. Actions in $D' \setminus \text{EDS}$ require escalation; actions in $D \setminus D'$ are prohibited regardless of authorization. This instantiates levels-of-autonomy governance [28] as an enforceable institution rather than a design guideline and makes Article 14 oversight operationally sustainable at scale.

The five-level taxonomy in Table 2 is an executive simplification, not a complete ordinal risk scale. Operationally, authority should be represented as an authority profile: execution autonomy, decision scope, materiality, reversibility, oversight latency, external exposure, data sensitivity, and tool criticality. Evidence of compliant operation may support a formally approved authority upgrade, but no agent earns authority automatically. Upgrades require documented human approval, independent validation where material, adversarial or stress evidence, stable domain assumptions, rollback controls, time limits, and recertification. Violations, population-level drift, or boundary pressure can trigger immediate downgrade or suspension.

4.5 C4: Runtime enforcement and containment

Training-time assurance does not survive deployment incentives [21, 22]; runtime governance supplies continuous verification with intervention capability. Rather than reducing heterogeneous risk to a single compensatory score, financial implementations should maintain a multidimensional *agent risk profile*: autonomy, adaptability, continuity, materiality, tool criticality, regulatory sensitivity, external exposure, security risk, reversibility, and protected-attribute sensitivity. A dashboard may optionally summarize a non-compensatory subset of that profile, but materiality and hard-gate dimensions must remain explicit: payments access, regulated customer impact, security exposure, and irreversible action rights should not be offset by low values elsewhere. Drift is detected through population-level monitors: behavioral-signature monitoring against baseline decision distributions, boundary monitoring that detects repeated approach to the EDS boundary before violation, cross-agent consistency monitoring that flags divergence in how similar contexts are governed, and dependency-graph monitoring over delegation, communication, shared state, shared models, shared data, shared tools, and common

Table 2: The ARIA five-level autonomy taxonomy as an executive view of a richer authority profile. Upgrades require formal approval and independent evidence where material; violations or population-level drift trigger mandatory review, downgrade, or suspension.

Level	Name	Operating condition
L0	Supervised	All outputs require human review before execution; novel or high-stakes decisions without constitutional precedent.
L1	Assisted	Agent proposes with full reasoning trace; human approves within defined windows; above materiality thresholds.
L2	Monitored	Autonomous execution within tight boundaries; real-time monitoring; automatic escalation on boundary contact.
L3	Delegated	Autonomous within broad boundaries; periodic human audit; stable domains with extensive compliance history.
L4	Autonomous	Post-hoc audit and exception-based review; fully characterized, low-catastrophic-risk domains only.

feedback signals. Signals feed policy decision points and policy enforcement points (PDPs/PEPs) at the orchestrator, tool gateway, API gateway, transaction layer, model wrapper, event bus, or human-approval service. Graduated containment moves from monitoring escalation to authority downgrade, sandboxing, and controlled shutdown. Each step is logged, reversible where appropriate, and designed to support post-deployment monitoring, supervisory evidence, trajectory-level model-risk evidence [16], and Article 72 post-market monitoring where the institution acts as provider.

4.6 C5: Adaptive policy learning

Static policy specifications are brittle under adversarial pressure, regulatory change, and distribution shift. This capability treats red-team findings, regulatory feedback, operational failures, and near misses as inputs to governed policy change. Red-team agent populations probe boundaries and feed failures into pre-registered amendment hypotheses; incidents are taxonomized; policy variants are evaluated offline, in sandbox, by replay, or on synthetic and controlled historical data; and updates are accepted only after evidence, owner approval, regression checks, and rollback planning. Antifragility remains a research hypothesis for this capability, not a demonstrated property of the architecture. The capability is a change-management and stress-testing process for value specifications, not an analogue of capital adequacy.

4.7 C6: Human oversight competence

Every human-in-the-loop obligation presupposes humans who can exercise meaningful oversight. Automation erodes exactly that capability, a problem long described as the ironies of automation and as automation misuse, disuse, and abuse [29, 30]. A credit analyst who has not exercised judgment for years cannot meaningfully review a credit agent. The required controls are deliberately narrow: an oversight-competence baseline before automation, periodic calibration exercises on realistic cases, and minimum practice plus recertification for accountable oversight personnel. Broader Protocol for Human Preservation in AI-Optimized Organizations (PHP-AIO) concerns about tacit-knowledge preservation and socio-institutional capital motivate the control objective [31], but the minimum architectural requirement is evidence that human oversight remains competent rather than ceremonial.

4.8 Mapping failure modes to controls

After introducing the six capabilities, Table 3 maps the non-compositionality taxonomy to observable signals and the primary ARIA controls most directly implicated by each failure mode.

5 Operationalizing ARIA in Financial Institutions

5.1 Regulatory evidence and supervisory questions

Table 4 gives the mapping in a practical form: the supervisory question each capability helps answer, the evidence it should leave behind, and the accountability pattern around it. The mapping turns the architecture into a candidate control set for pilot implementation, assurance testing, and supervisory assessment. The claims are intentionally prudential: the controls may help operationalize evidence needs under existing regimes, but the table is not a legal interpretation. The ownership model follows the three lines of defence and a responsible-accountable-consulted-informed (RACI) pattern: the first line owns agents, authority, operation,

Table 3: How the taxonomy maps to observable signals and primary ARIA controls. The mapping is illustrative: deployments may require multiple controls and domain-specific evidence.

Type	Example	Observable signal	Primary capability
Outcome	Approval disparity	Population gap or tail outcome	C2/C4
Interaction-induced	Collusion or strategic gaming	Trajectories, prices, coordinated timing	C2/C4/C5
Shared-environment	Common-signal concentration	Correlated dependency or exposure shift	C2
Cross-agent information	Shared-state leakage	Graph, memory, or state telemetry	C2/C4
Control-plane	Delegation chain bypass	Permission path or escalation trace	C3/C4
Accountability	Outcome without owner	Missing lineage, responsibility assignment, or audit trail	C1/C6

Table 4: Simplified regulatory mapping for ARIA. Each row states the supervisory question, compact regulatory anchors, the evidence produced, and a responsible-accountable-consulted-informed (RACI) accountability pattern. Article numbers refer to the EU AI Act [2]; U.S. model-risk references to SR 26-2 and historical SR 11-7 [3, 8]; UK and EU anchors to FCA Consumer Duty, PRA SS1/23, and DORA [9, 10, 14]; and agent-governance anchors to MAS FEAT, IMDA, and NIST [11–13].

C	Supervisory question and anchors	Evidence produced	RACI pattern
C1	Policy specification. Are values, fair-lending rules, and jurisdictional mappings specified consistently? AI Act Arts. 9/17; ECOA/Reg B; FEAT.	Policy ontology; jurisdictional mappings; exception and amendment trail.	A: board and senior mgmt.; R: responsible AI, compliance, legal, business risk; internal audit: third line
C2	Population assurance. Does observed population behavior remain concordant with expected behavior documented in the deployment reference model? AI Act Art. 26 and, for provider roles, Art. 72; SR 26-2; FCA Consumer Duty; PRA SS1/23.	M_2 observed-versus-expected time series; calibrated thresholds; dependency graph; deployment reference model.	R: first-line telemetry; C: model risk and independent risk; A: chief risk officer or risk committee; internal audit: third line
C3	BDA. Are agents acting within authorized profiles and escalation limits? AI Act Arts. 14/26; IMDA bounded autonomy; NIST authorization levels.	Authority profile; escalation logs; approved upgrade and downgrade history.	A/R: business; C: risk, compliance, legal; internal audit: third line
C4	Runtime enforcement. Can the institution detect drift and intervene before harm materializes? AI Act Art. 26 and, for provider roles, Art. 72; SR 26-2; FEAT; IMDA/NIST runtime monitoring.	PDP/PEP logs; drift logs; containment register; reversal records.	R: platform/control; A: business and risk control; C: compliance and security; internal audit: third line
C5	Adaptive policy learning. Are incidents and stress tests converted into governed, regression-tested policy updates? AI Act Art. 9; supervisory stress testing; DORA digital-operational-resilience logic.	Red-team register; amendment hypotheses; sandbox/replay results; rollback records.	R: first-line control or policy owner; A: risk committee; C: model risk, business, legal, compliance; internal audit: audit only
C6	Human oversight competence. Are humans still competent to exercise meaningful oversight? AI Act Arts. 14/4; senior manager accountability regimes.	Competence baselines; calibration exercises; recertification records.	A/R: senior managers and human-resources learning; C: risk and compliance; internal audit: third line

and customer outcomes; the second line sets risk policy, challenges metrics, and oversees containment; the third line independently audits control design and effectiveness; board and senior management retain accountability.

5.2 Minimum Viable ARIA

The six-capability architecture does not have to arrive as a single enterprise transformation. A minimum viable implementation is an auditable control bundle for any financial institution deploying agent populations in high-impact workflows. It has nine elements:

1. An agent registry recording identity, version, owner, purpose, model lineage, and deployment context.
2. An authority matrix specifying permitted actions by agent, workflow, tool, data domain, and jurisdiction.
3. An interaction and dependency graph recording delegation, communication, shared-state, shared-model, shared-data, shared-tool, and common-feedback dependencies.
4. A constitutional core translated into workflow-level operating rules, including escalation thresholds for collective outcomes.
5. Population-level monitors for fairness, conduct, and other outcome distributions, not only per-agent constraint checks.
6. Runtime telemetry over tool calls, delegated decisions, boundary contacts, and cross-agent state sharing.
7. A graduated containment policy linking monitor signals to monitoring escalation, autonomy downgrade, sandboxing, or shutdown.

8. An incident-to-policy-update workflow that turns violations, near misses, and red-team findings into versioned constitutional amendments.
9. Periodic human-calibration exercises that preserve the expertise required for meaningful oversight.

This bundle does not implement every feature of ARIA, but it makes identity, authority, dependency topology, collective outcomes, intervention rights, and human competence examinable from the first deployment. Institutions can then add richer deliberation, richer M_2 reporting, and adaptive policy-learning evidence as maturity accumulates.

5.3 Accountability and auditability

A governance architecture is only as good as the accountability it produces. Two properties matter for fairness and accountability: who answers for what, and whether outsiders can verify it.

Responsibility assignment. Non-compositionality creates responsibility gaps: collective violations for which no individual agent, and often no individual human, is answerable [24]. ARIA narrows that gap not by verifying gap-freeness, which is intractable for the reasons given in Section 3, but by constructing answerable roles for each capability (Table 4). The first line owns agents, authority decisions, workflow operation, and customer outcomes; the second line sets risk policy, challenges metrics, and oversees containment; the third line independently audits control design and effectiveness; and board and senior management retain accountability for risk appetite, policy approval, and oversight competence. Internal audit is not an operating owner of adaptive policy change. Each collective outcome should trace to a capability, and each capability to a RACI pattern and named human role, following the logic of senior-manager accountability regimes. The architecture is therefore an accountability allocation, not only a control system.

Auditability. Every capability emits a designated artifact (Table 4): versioned policy specifications with deliberation records; M_2 time series; authorization matrices and escalation logs; containment registers; red-team and amendment registers; and expertise-maintenance records. These artifacts support internal audit testing, supervisory assessment, and aggregate or redacted external accountability. The design goal is simple: the question “was the institution governing its agent population?” should become an evidence question, not a trust question.

Power and contestation. Federated policy formation redistributes specification power from a single engineering function to a structured multi-stakeholder process. The policy-specification process creates channels for affected-community contestation and supervisory scrutiny, while accountable institutional owners retain amendment authority and pre-registration prevents silent revision. This does not settle the deeper political question of whose values should govern. It makes the allocation explicit and contestable, which is the precondition for asking that question in a serious way.

6 Conceptual Demonstrations

6.1 Demonstration design and reconstruction parameters

The demonstrations are deliberately narrow. They are not benchmarks, production effect-size estimates, or evidence that ARIA is validated. They illustrate two mechanisms that a future field program must test: (E1) local equal-treatment checks can coexist with aggregate disparity under shared-signal and measurement-quality coupling, and population-level containment can reduce that disparity in the stylized setting; (E2) in a deliberately constructed two-stage drift regime, an observed-versus-expected M_2 monitor alarms earlier than a prohibited-outcome monitor. Table 5 reports reconstruction-level parameters, seeds, monitoring rules, and evaluation metrics; Section 9 states the limits.

6.2 Shared-signal thin-file exclusion and population-level containment

Setup. $N = 200$ credit components evaluate applicants from two groups with identical true creditworthiness. Group G_1 is thin-file, with estimation noise $\sigma_1 = 0.65$ versus $\sigma_0 = 0.30$, matching the structure of historically excluded populations with sparse credit records. Each component satisfies a local equal-treatment check conditional on its own risk estimate ($\kappa_i = 1$ throughout). Components adapt a variance-aversion parameter through a shared portfolio-loss signal; this is a shared-environment coupling mechanism, not evidence of strategic coordination. The collective principle bounds population approval-rate disparity at $\delta = 0.05$, a simulation parameter rather than a legal threshold. We compare local controls alone with ARIA-style population monitoring and graduated containment, using 20 seeds and $T = 400$.

Results (Figure 3). Without population-level governance, collective disparity grows from 0.05 to 0.41 in this stylized setting, or $8\times$ the chosen bound, with zero local-check violations. This does not show that all

Table 5: Methodological specification of the two simulations. The table gives reconstruction-level parameters, not production calibration values.

Experiment	Scale	Key settings	Metric
Shared-signal thin-file exclusion	$N = 200, T = 400, 20$ seeds	Equal true creditworthiness ($\mu = 0.5$); estimation noise $\sigma_0 = 0.30, \sigma_1 = 0.65$; collective bound $\delta = 0.05$; variance aversion initialized as $ \mathcal{N}(0.3, 0.05) $ and updated by clipped $\mathcal{N}(0.012, 0.004)$; containment downgrades the top decile when gap $> \delta$ and applies corrective feedback proportional to overshoot.	Approval-rate gap, individual local-check violations, 10–90% band.
Two-stage behavior drift	$T = 300, K = 8, 200$ episodes	Category 0 prohibited; clean and drift episodes alternate; 400 sampled decisions per step; expected-behavior reference calibrated from a Dirichlet baseline with prohibited mass 0.01; drift starts uniformly in steps 60–139; phase 1 permutes benign mass for 70 steps; phase 2 targets prohibited mass 0.35; violation threshold 0.10; detector thresholds matched at false-positive rate $\leq 5\%$.	Lead time before observable violation, receiver-operating-characteristic area.

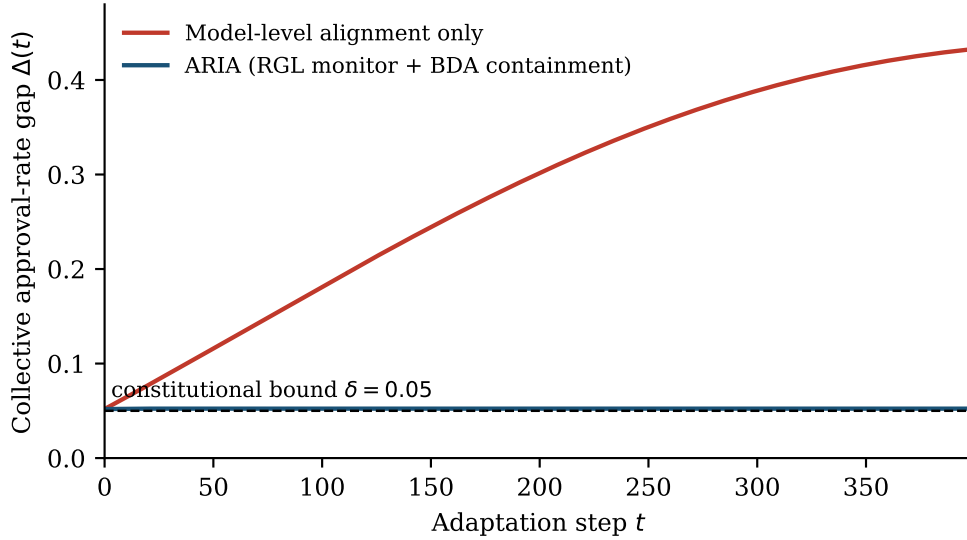


Figure 3: Shared-signal thin-file exclusion in a stylized credit population (mean, 10–90% band, 20 seeds). Local controls alone: collective disparity reaches $8\times$ the chosen bound $\delta = 0.05$ with zero local-check violations. Population-level containment holds disparity close to the bound in the stylized setting.

procedurally consistent components become collectively discriminatory. It shows that a local equal-treatment check conditional on imperfect, group-dependent measurements can be incompatible with an aggregate disparity constraint. Under population-level containment, the same population under the same pressure is held at 0.052, within 4% of the bound, through authority downgrades and corrective feedback. The containment rule is intentionally idealized and has direct access to the monitored gap; implementation delays, noisy group labels, and utility trade-offs are deferred to field validation. The demonstration motivates the control problem; it does not estimate real credit effects or the business cost of containment.

6.3 M_2 as an observed-versus-expected behavior indicator

Setup. Episodes last $T = 300$ steps over $K = 8$ action categories, with category 0 prohibited. Drift is deliberately constructed in two phases: first probability mass reorganizes among benign categories, and only later does the prohibited category cross the violation threshold. We compare two detectors with the same sample budget, 400 decisions per step, and matched false-positive rate ($\leq 5\%$ on clean episodes): an outcome monitor watching prohibited-action frequency and the M_2 observed-versus-expected behavior indicator of Eq. (1). We run 200 evaluation episodes, alternating clean and drifted cases.

Results (Figure 4). Both detectors eventually discriminate perfectly in this constructed regime, so the

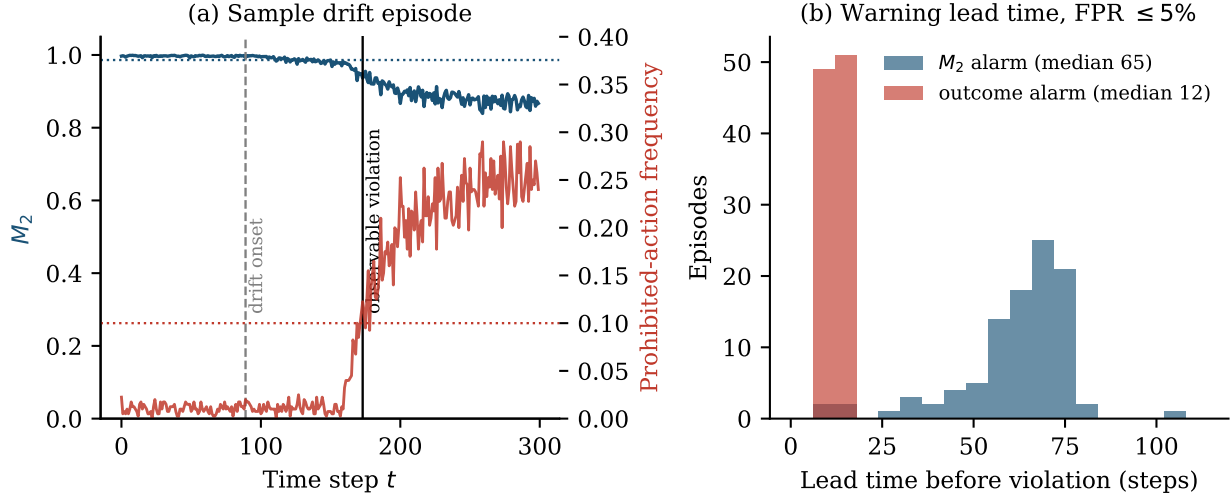


Figure 4: Observed-versus-expected drift warning. (a) A constructed two-stage drift episode: M_2 degrades as observed behavior diverges from the expected distribution during benign reorganization; prohibited-action frequency rises only later. (b) Lead time before observable violation at matched false-positive rate $\leq 5\%$: median 65 steps (M_2) vs. 12 (outcome monitoring), across the drifted evaluation episodes.

result should not be read as a general superiority claim. The difference is timing under this data generator: the observed-versus-expected distributional monitor alarms during benign-category reorganization, whereas the prohibited-outcome monitor cannot alarm until mass reaches the prohibited category. The median lead times are 65 and 12 steps, respectively, at matched false-positive rate. The demonstration therefore supports a narrower claim: when drift first appears in monitored distributional structure before an observable violation, an observed-versus-expected statistic can provide earlier warning than a univariate prohibited-outcome monitor.

7 Field Validation and Standardization Path

7.1 Field validation methodology

Validating ARIA beyond simulation requires an environment in which organizational alignment can actually fail. Global financial institutions operating across multiple jurisdictions are a natural setting. They combine *regulatory complexity*, including simultaneous EU AI Act, SR 26-2, FCA Consumer Duty, PRA SS1/23, MAS, DORA, and analogous requirements across multiple jurisdictions, which is the conflict F-CAI addresses and which the Financial Stability Board identifies as an active governance challenge [32]; *high-precision measurability* of credit, fraud, and compliance outcomes; *heterogeneous agent populations* spanning model families, training histories, and mandates; and *continuous adversarial pressure*, which supplies the stress conditions that adaptive policy learning requires.

7.2 Five-stage validation program

We propose a five-stage program, each stage a pre-registered empirical question mapped to a capability, and each falsifiable:

Stage 1 (policy specification). Do structured participatory policy specifications achieve higher measured adherence across heterogeneous stakeholder populations than expert-authored ones? Comparative behavioral testing on identical agent populations; preference elicitation from supervisory, employee, and affected-customer stakeholder groups.

Stage 2 (M_2). Does M_2 detect policy or behavioral drift by comparing observed population behavior with expected behavior before observable violations in production populations, and does it outperform simpler baselines such as multivariate cumulative-sum tests, Page-Hinkley tests, population-stability indices, classifier-based two-sample tests, embedding drift, transition-matrix drift, and multivariate outcome monitors? Longitudinal M_2 tracking against recorded violation events, with receiver-operating-characteristic analysis, lead-time, ablation, and calibration analysis, is the stage that would establish or refute M_2 as a useful reportable indicator.

Stage 3 (BDA). Does bounded decision authority reduce violation frequency and severity at acceptable cost? Matched-domain A/B comparison and fault-injection testing should measure violation rate, violation duration, maximum overshoot, containment latency, mediated-action coverage, false-containment rate, business utility loss, customer delay, and operational workload.

Stage 4 (adaptive policy learning). Do governed incident-to-policy amendments improve future violation metrics without unacceptable regression, complexity growth, or overfitting to the last incident? Longitudinal tracking should compare static, reactive, and adaptive amendment processes against adaptive red-team exercises and should measure improvement, regression on prior scenarios, policy complexity, owner approval latency, rollback frequency, and implementation cost.

Stage 5 (human oversight competence). Does mandated expertise maintenance improve measured human oversight quality at 6-, 12- and 24-month horizons against control organizations?

The program is designed as a collaborative agenda with pre-registration of both hypotheses and policy changes. We note the obvious conflict-of-interest hazard: an institution evaluating its own governance architecture has an incentive to find that it works. Pre-registration, external academic partners with publication rights independent of results, and regulatory access to the raw artifacts are the mitigations we consider minimal.

7.3 Toward a supervisory reference architecture

SR 11-7 did not invent model validation. It supplied a historical shared vocabulary and control taxonomy that made model risk governable, assessable, and comparable across institutions. SR 26-2 now provides the revised U.S. supervisory reference point, superseding and replacing SR 11-7 while preserving the need for risk-based governance, validation, monitoring, and controls [3, 8]. Multi-agent AI in regulated industries may be at a similar early stage. Supervisory evidence needs increasingly concern collective behavior (Section 2), while coupled agentic workflows still lack a shared governance architecture. ARIA is positioned as a candidate research architecture for that role, not as a mature supervisory standard. The claim is deliberately falsifiable: the architecture specifies candidate statistics, named artifacts, and named accountable roles, so adoption, assessment, and cross-institution comparison can become concrete research questions.

Three properties would make a governance framework standardizable if validated. *Regulatory anchoring:* each control is connected to evidence needs institutions already face (Table 4). *Capability separability:* an institution could pilot elements incrementally, for example runtime monitoring first, bounded authority second, and federated constitutional formation as maturity grows. *Supervisability:* the artifacts support assurance and the metrics support trend analysis. The validation program that could substantiate the architecture consists of staged field studies in production-like or production agentic workflows with pre-registered hypotheses, negative-results publication rights, and independent scrutiny. Executing that program is the agenda this paper opens.

8 Related Work

8.1 Multi-agent alignment and collusion

Cooperative-equilibrium constructions [25] show that collective alignment can be achieved under designed incentives. ARIA supplies the institutional machinery that makes such incentives durable, updatable, and legitimate under regulatory change. Empirical collusion results [4, 5] ground Proposition 1 in observed behavior rather than theoretical possibility alone. ARIA differs by targeting regulated-finance outcomes, fair-lending non-compositionality, accountable human functions, and audit artifacts rather than collusion control alone.

8.2 Agentic finance and financial supervision

Agentic-finance and financial-services runtime-governance work describes a shift from isolated prediction models toward autonomous systems, trajectory telemetry, continuous authorization, temporal conformance checking, and tiered containment [15, 16]. ARIA absorbs this trajectory-level insight but emphasizes population-level fairness constraints, observed-versus-expected behavior monitoring (M_2), constitutional amendment, and human-competence preservation as finance-specific governance capabilities.

8.3 Fairness in coupled and multi-agent systems

Recent work on financial multi-agent bias and fairness in agentic AI treats fairness as a dynamic, interactional, and system-level property [6, 7]. The thin-file disparity demonstration in Section 6.2 should be read against that literature and against classic concerns about measurement quality, group-dependent uncertainty, selective labels, proxy discrimination, and calibration-fairness trade-offs, rather than as a claim that all fairness failures are uniquely multi-agent.

8.4 Organizational cybernetics and systems safety

The population-assurance capability builds on Beer’s VSM [27]. We extend that line to agent-population governance and use M_2 as a candidate continuous observed-versus-expected behavior statistic rather than as a complete model of organizational self-knowledge. ARIA also inherits from older safety ideas: runtime assurance and Simplex-style separation [33], system-theoretic accident modeling [34], defense in depth, common-mode failure analysis, and high-reliability organizing. Its novelty is not that monitoring and containment are new, but that they are organized for adaptive agentic workflows under financial accountability.

8.5 Democratic and value-sensitive alignment

Gabriel [26] supplies the normative foundation, and Collective Constitutional AI [18] provides the main empirical anchor for participatory formation. ARIA extends that logic from model values to continuous organizational policy formation.

8.6 Runtime governance, identity and agent safety

MI9 [35] provides a published runtime-governance architecture and an Agency-Risk Index that motivates non-compensatory summaries of agent risk profiles. Policies-on-paths formalize compliance as a function of agent identity, partial execution path, proposed action, and organizational state [36]. Agent identity infrastructure makes actions governable through delegated authorization and accountability for nondeterministic agents [37]. Agent misalignment benchmarking [22] and alignment faking [21] establish the training-deployment gap runtime enforcement exists to close. Responsibility-gap complexity [23, 24] motivates governance that does not depend on exhaustive verification.

8.7 Constitutional AI and its extensions

CAI [17] makes model values explicit and auditable. ARIA is not an alternative to this line of work but the organizational layer above it: it presupposes individually governed agents and addresses what they do collectively.

8.8 Adaptive governance and antifragility

de la Chica et al. [38] study stress as a detectable signal for antifragility-compatible regimes in multi-agent LLM systems, while Taleb [39] supplies the design philosophy behind the adaptive-policy research agenda. In this manuscript, however, antifragility is treated as a hypothesis for future validation, not as an empirical result.

9 Limitations

Our claims are bounded in five ways. *The formal result is only a counterexample*: Proposition 1 shows that local checks do not logically imply collective compliance, not that every agent population will fail or that strategic equilibria are always the mechanism. *The design hypothesis is operational*: it requires mediated actions, monitorable collective constraints, bounded response latency, and acceptable trade-offs in false containment, delay, and business utility. Principles that resist statistical operationalization, such as dignity or non-manipulation in the thick sense, fall outside it, and characterizing that boundary remains open. *The simulations are mechanism illustrations*: agents are analytic decision rules rather than LLMs; the thin-file disparity result depends on shared-signal and measurement-quality coupling; the drift regime is deliberately two-stage; both detectors in Experiment 2 achieve ceiling discrimination, so only the lead-time contrast is informative. Effect sizes will not transfer to production; mechanism structure is the claim. *M_2 is support-dependent*: because it compares observed and expected behavior over a chosen support, it can miss harms that preserve monitored marginals, reflect benign distributional change, or fail under a miscalibrated reference model. *Governance can be theater*: an institution can adopt the artifacts without the substance. Independent assessment of the artifacts, independent validation, and publication of negative results are the backstops.

10 Conclusion

Fairness and safety controls attached to individual models need not see, and therefore may fail to prevent, collective failures produced by coupled agentic workflows. We have given a finite counterexample, two transparent mechanism demonstrations, and a six-capability governance architecture mapped to evidence needs across multiple regulatory regimes. The architecture also specifies the accountability allocation and audit artifacts that make such governance assessable. The blind spot is structural; the remedy proposed here is architectural and still requires field validation. Production evidence, cross-institutional replication, operational specification, and regulatory engagement would be required before ARIA could be considered a reference architecture

in any stronger sense. We therefore specify a falsifiable five-stage program (Section 7) rather than a claim of validation. If M_2 does not add value over simpler drift baselines in real agent populations, if bounded authority does not reduce violation severity at acceptable cost, or if governed policy updates do not improve future outcomes without regression and complexity burden, the architecture should be revised or narrowed on that evidence. The resulting agenda is a falsifiable validation program, not a declaration of effectiveness.

Declarations

Ethical considerations

This work analyzes governance failures that could, if detailed for a specific deployment, inform evasion of the very controls proposed; we therefore describe detection mechanisms at the architectural level and withhold deployment-specific thresholds. The simulations use synthetic data only; no human subjects or personal data were involved. The thin-file disparity demonstration models a redlining-like exclusion pattern to expose a control failure, not to provide a method; the containment mechanism is presented alongside it. Field validation will require ethics review, regulator engagement, and protections for affected customer populations.

Adverse impacts

Two adverse-impact channels merit statement. First, a population-level governance architecture, if adopted as compliance theater, could *legitimize* multi-agent deployments without delivering substantive protection. We mitigate that risk by designing every capability around independently assessable artifacts and by naming accountable human roles, but the risk is real and supervision-dependent. Second, the same monitoring machinery that detects collective discrimination could be repurposed for worker or customer surveillance. The architecture’s policy-specification capability is the intended constraint, and repurposing risk applies explicitly in deployment and supervision contexts. We judge the counterfactual, ungoverned agent populations in high-stakes finance, to carry substantially greater expected harm.

Transparency statement

Generative AI tools were used to support literature triage, copy-editing, technical consistency checks, and compilation-log inspection. The authors remain responsible for all claims, references, analysis, and final wording.

Acknowledgments

This work forms part of a broader research agenda on agent-population governance in regulated finance.

References

- [1] United States Congress. Equal Credit Opportunity Act, 15 U.S.C. § 1691 et seq., and Regulation B, 12 C.F.R. Part 1002, 1974.
- [2] European Parliament and Council. Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (ai act), 2024. Official Journal of the European Union, L series.
- [3] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency and Federal Deposit Insurance Corporation. Revised Guidance on Model Risk Management (SR 26-2). Technical report, Board of Governors of the Federal Reserve System, 2026. URL <https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm>. April 17, 2026. Supersedes and replaces SR letter 11-7.
- [4] Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. Algorithmic collusion by large language models. *arXiv preprint arXiv:2404.00806*, 2024.
- [5] Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip H. S. Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among AI agents: Multi-agent deception via steganography. *Advances in Neural Information Processing Systems (NeurIPS)*; *arXiv:2402.07510*, 2024.
- [6] Maeve Madigan, Parameswaran Kamalaruban, Glenn Moynihan, Tom Kempton, David Sutton, and Stuart Burrell. Emergent bias and fairness in multi-agent decision systems. *arXiv preprint arXiv:2512.16433*, 2026. doi: 10.48550/arXiv.2512.16433. URL <https://arxiv.org/abs/2512.16433>.
- [7] Rajesh Ranjan, Shailja Gupta, and Surya Narayan Singh. Fairness in agentic AI: A unified framework for ethical and equitable multi-agent system. *arXiv preprint arXiv:2502.07254*, 2025. doi: 10.48550/arXiv.2502.07254. URL <https://arxiv.org/abs/2502.07254>.

- [8] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency. Supervisory guidance on model risk management (SR 11-7). Technical report, Federal Reserve, 2011.
- [9] Financial Conduct Authority. Final non-handbook guidance for firms on the Consumer Duty (FG22/5). Technical report, Financial Conduct Authority, July 2022. URL <https://www.fca.org.uk/publication/finalised-guidance/fg22-5.pdf>.
- [10] Prudential Regulation Authority. Model Risk Management Principles for Banks (SS1/23). Technical report, Bank of England, May 2023. URL <https://www.bankofengland.co.uk/prudential-regulation/publication/2023/may/model-risk-management-principles-for-banks-ss>.
- [11] Monetary Authority of Singapore. Principles to promote fairness, ethics, accountability and transparency (FEAT) in the use of AI and data analytics in singapore’s financial sector. Technical report, MAS, 2018.
- [12] Infocomm Media Development Authority. Model AI governance framework for agentic AI. Technical report, IMDA, Singapore, 2026. URL <https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf>.
- [13] National Institute of Standards and Technology. AI agent standards initiative. Technical report, NIST, Gaithersburg, MD, 2026. URL <https://www.nist.gov/artificial-intelligence/ai-agent-standards-initiative>.
- [14] European Parliament and Council. Regulation (EU) 2022/2554 on digital operational resilience for the financial sector, 2022. URL <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R2554>. Official Journal of the European Union, L 333/1.
- [15] Irene Aldridge, Jolie An, Riley Burke, Michael Cao, Chia-Yi Chien, Kexin Deng, Ruipeng Deng, Yichen Gao, Olivia Guo, Shunran He, Zheng Li, George Lin, Weihang Lin, Percy Lyu, Alex Ng, Qi Wang, Hanxi Xiao, Dora Xu, Yuanyuan Xue, Sheng Zhang, Sirui Zhang, Yun Zhang, Sirui Zhao, Xiaolong Zhao, Yihan Zhao, and Waner Zheng. Agentic artificial intelligence in finance: A comprehensive survey. *arXiv preprint arXiv:2604.21672*, 2026. doi: 10.48550/arXiv.2604.21672. URL <https://arxiv.org/abs/2604.21672>.
- [16] Lukasz Szpruch, Agus Sudjianto, Tanveer Bhatti, and Gary Ang. Scalable runtime governance for agentic AI in financial services. Technical report, SSRN preprint, 2026. URL <https://ssrn.com/abstract=6567199>. April 13, 2026. DOI: 10.2139/ssrn.6567199.
- [17] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- [18] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective constitutional AI: Aligning a language model with public input. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.
- [19] Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, et al. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*, 2024.
- [20] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.
- [21] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.
- [22] Akshat Naik, Patrick Quinn, Guillermo Bosch, Emma Gouné, Francisco Javier Campos Zabala, Jason Ross Brown, and Edward James Young. AgentMisalignment: Measuring the propensity for misaligned behaviour in LLM-based agents. *arXiv preprint arXiv:2506.04018*, 2025. doi: 10.48550/arXiv.2506.04018. URL <https://arxiv.org/abs/2506.04018>.
- [23] Junli Jiang and Pavel Naumov. Responsibility gap and diffusion in sequential decision-making mechanisms. *arXiv preprint arXiv:2507.02582*, 2025.
- [24] Filippo Santoni de Sio and Giulio Mecacci. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, 34(4):1057–1084, 2021.
- [25] Yunhao Liang et al. Everyone contributes! incentivizing strategic cooperation in multi-LLM systems via sequential public goods games. *arXiv preprint arXiv:2508.02076*, 2025. URL <https://arxiv.org/abs/2508.02076>.

- [26] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, 2020.
- [27] Stafford Beer. *Brain of the Firm*. Allen Lane, London, 1972.
- [28] K. J. Kevin Feng, David W. McDonald, and Amy X. Zhang. Levels of autonomy for AI agents. *arXiv preprint arXiv:2506.12469*, 2025. doi: 10.48550/arXiv.2506.12469. URL <https://arxiv.org/abs/2506.12469>.
- [29] Lisanne Bainbridge. Ironies of automation. *Automatica*, 19(6):775–779, 1983. doi: 10.1016/0005-1098(83)90046-8.
- [30] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2): 230–253, 1997. doi: 10.1518/001872097778543886.
- [31] Jose Manuel de la Chica Rodriguez, Jairo Rodriguez Arias, and Spyridon Chouliaras. When Not to Automate: A formal protocol for human preservation in AI-optimized organizations, 2026. URL <https://arxiv.org/abs/2607.15944>. arXiv:2607.15944.
- [32] Financial Stability Board. The financial stability implications of artificial intelligence. Technical report, Financial Stability Board, Basel, 2024.
- [33] Lui Sha, Ragunathan Rajkumar, and Michael H. Gagliardi. The Simplex architecture: Using runtime monitoring to provide dependable system upgrade. *Proceedings of the IEEE Real-Time Systems Symposium*, pages 413–414, 1998.
- [34] Nancy G. Leveson. *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press, Cambridge, MA, 2011.
- [35] Charles L. Wang, Trisha Singhal, Ameya Kelkar, and Jason Tuo. MI9 – agent intelligence protocol: Runtime governance for agentic AI systems. *arXiv preprint arXiv:2508.03858*, 2025. doi: 10.48550/arXiv.2508.03858. URL <https://arxiv.org/abs/2508.03858>.
- [36] Maurits Kaptein, Vassilis-Javed Khan, and Andriy Podstavnychy. Runtime governance for AI agents: Policies on paths. *arXiv preprint arXiv:2603.16586*, 2026. doi: 10.48550/arXiv.2603.16586. URL <https://arxiv.org/abs/2603.16586>.
- [37] Paras Singla. Agent identity protocol (AIP): Decentralized identity and delegation for AI agents. Internet-Draft draft-singla-agent-identity-protocol-03, Internet Engineering Task Force, June 2026. URL <https://datatracker.ietf.org/doc/draft-singla-agent-identity-protocol/03/>. Work in Progress.
- [38] Jose Manuel de la Chica, Juan Manuel Vera, and Jairo Rodríguez. When Stress Becomes Signal: Detecting antifragility-compatible regimes in multi-agent LLM systems, 2026. URL <https://arxiv.org/abs/2605.02463>. arXiv:2605.02463.
- [39] Nassim Nicholas Taleb. *Antifragile: Things That Gain from Disorder*. Random House, New York, 2012.