



Original Article

Agent Readiness Scoring: A Capability Taxonomy for Onboarding Enterprise AI Agents into Regulated Operations

Darshan Bhansali

Principal System Architect, AT&T - Consumer Technology Experience, Dallas, TX.

Received On: 28/04/2025

Revised On: 08/05/2025

Accepted On: 23/05/2025

Published on: 13/06/2025

Abstract - As enterprises move from single-assistant deployments to agent fleets, the binding constraint shifts from model quality to onboarding governance: deciding which agent may act, on what surface, under what oversight. Existing work optimizes individual agent architecture and benchmark task success, and existing enterprise controls reduce to static role allowlists or committee review, neither of which expresses the property that actually determines exposure, namely whether an erroneous action can be undone. This paper presents AGORA (Agent Governance, Onboarding, and Readiness Assessment), a framework developed during evaluation and enablement of approximately 65 AI agents exposing more than 100 distinct capabilities inside a Fortune 50 telecommunications environment. AGORA decomposes every agent capability into a four-dimensional signature over action scope, data sensitivity, effect reversibility, and human checkpoint requirement, and organizes signatures as a partially ordered lattice. A monotone map from this lattice to a graduated autonomy ladder yields the Agent Readiness Score and the permitted deployment tier, so that autonomy is granted per capability rather than per agent. Promotion between tiers is adjudicated by a sequential probability ratio test over observed action outcomes, replacing calendar-based or approval-based gating with an evidence-based stopping rule. Fleet-level redundancy is detected through Jaccard similarity over capability signatures with community detection on the resulting graph. Silent scope expansion after model updates is detected by Kullback-Leibler divergence between declared and observed tool-invocation distributions. Against manual review, static role-based control, calendar-staged rollout, and an LLM-judge risk classifier, AGORA achieved 0.934 precision and 0.958 recall in restricting capabilities warranting restriction, reduced unauthorized action attempts by 91.3 percent, and detected post-update scope expansion with a median lead time of 3.1 days.

Keywords - Agent Governance, Capability Taxonomy, Graduated Autonomy, Reversibility Risk, Enterprise AI Onboarding, Scope Drift Detection, Sequential Testing.

1. Introduction

1.1. Background

Enterprise AI deployment has shifted from a single conversational assistant to fleets of task-specialized agents that invoke tools, query systems of record, and mutate operational state. In a large telecommunications environment, dozens of agents may expose overlapping capabilities across incident response, provisioning, customer care, and knowledge retrieval. Each agent introduces an authorization surface, an audit obligation, and a failure mode. The governing question is no longer whether an agent performs well on a benchmark, but which of its capabilities may execute against which systems, with what human oversight, and under what evidentiary standard for expanding that permission.

1.2. Limitations of Existing Approaches

Static role-based access control assigns permissions to an agent identity and is blind to the distinction between a reversible read and an irreversible mutation performed with the same credential. Manual review boards produce defensible decisions but scale poorly and drift into approval-by-familiarity. Calendar-staged rollout advances agents on

elapsed time rather than observed evidence, promoting rarely exercised agents on the same schedule as heavily exercised ones. LLM-judge risk classifiers offer scale but yield non-reproducible verdicts unsuitable for regulatory audit. Emerging capability-catalogue practice records what an agent can do without recording what happens when it does so incorrectly.

1.3. Proposed Solution and Contribution Summary

We contribute a four-dimensional capability signature in which reversibility and data sensitivity are first-class dimensions, and we organize signatures as a partially ordered lattice so that autonomy is granted per capability rather than per agent identity. We contribute a monotone lattice-to-ladder mapping producing the Agent Readiness Score, guaranteeing that no capability receives autonomy exceeding that of any capability dominating it. We contribute a sequential probability ratio test for tier promotion, giving a principled stopping rule with bounded error rates. We contribute fleet-level redundancy detection through signature similarity and community detection, and a divergence-based detector for silent scope expansion following model updates.

2. Related Work and Background

2.1. Conventional Approaches

Traditional enterprise integration governs non-human actors through service accounts, role assignments, and periodic entitlement recertification. Controls are identity-scoped and coarse. An agent granted a write role holds that role uniformly across every operation the role permits. Recertification occurs on fixed cycles and verifies role membership rather than behavioural conformance, so an actor whose effective behaviour has changed passes recertification unchanged.

2.2. Newer and Modern Approaches

Recent agent research emphasizes planning architectures, tool-use reliability, and multi-agent coordination, evaluated on task-completion benchmarks. Safety work concentrates on prompt-injection resistance, sandboxing, and output filtering. These contributions address whether an individual agent behaves correctly. They do not address which of a fleet's capabilities warrant execution rights in a production estate, nor how those rights should expand as evidence accumulates.

2.3. Related Hybrid and Alternative Models

Human-in-the-loop autonomy frameworks from robotics and aviation define levels of automation and are the closest antecedent to our ladder. Risk-tiered classification schemes from emerging regulatory practice group applications by domain and consequence. Both establish that autonomy should be graduated. Neither provides a computable per-capability score, an evidence-based promotion criterion, or a fleet-level view that detects redundancy and drift across many concurrently deployed agents.

2.4. Summary of Research Gap

The literature offers per-agent evaluation and abstract autonomy levels. It does not offer a reproducible, auditable mechanism that maps concrete capability properties to concrete execution rights, promotes on statistical evidence, and monitors a fleet for redundancy and post-update scope expansion. AGORA addresses that gap with a documented deployment across roughly 65 agents and over 100 capabilities.

3. Proposed Methodology

3.1. Overview

AGORA governs capabilities, not agents. Each declared capability receives a signature, a lattice position, a readiness score, and an autonomy tier. An agent's effective permission is the union of its capability grants.

3.2. The Capability Signature

Every capability is characterized on four ordinal dimensions. Action scope ranges from read-only retrieval through single-record mutation to bulk or cross-system

mutation. Data sensitivity ranges from public reference material through internal operational data to regulated customer and identity data. Reversibility ranges from trivially undoable, through undoable with a compensating transaction, to externally observable and therefore irreversible, such as a customer notification or a downstream provisioning call. Checkpoint requirement records whether a human confirmation is architecturally enforced, advisory, or absent.

3.3. The Capability Lattice

Signatures are ordered componentwise, forming a partially ordered set with meet and join. Capability x dominates y when x is at least as broad, sensitive, irreversible, and unsupervised on every dimension. Domination is the structural invariant of the framework: the autonomy assignment must be antitone with respect to it, so a dominating capability can never hold more autonomy than a capability it dominates.

3.4. The Agent Readiness Score

Readiness combines an evidence term and an exposure term. The evidence term is the posterior mean of observed action correctness under a Beta prior, penalized by the variance so that sparsely exercised capabilities cannot score highly on a small favourable sample. The exposure term is derived from the lattice position, with irreversibility weighted superlinearly because the cost of an irreversible error is not proportional to its frequency. The score is the evidence term discounted by exposure.

3.5. The Graduated Autonomy Ladder

Three tiers are defined. Advisory permits proposal output only, with no tool execution. Supervised permits execution behind a blocking human checkpoint. Bounded autonomous permits execution within declared parameter envelopes, with asynchronous review and mandatory compensating-transaction registration. The tier is assigned by thresholding the readiness score, subject to the antitonicity constraint from the lattice.

3.6. Evidence-Based Promotion

Promotion is adjudicated by a sequential probability ratio test on the stream of reviewed action outcomes, comparing a null hypothesis of unacceptable error rate against an alternative of acceptable error rate at fixed Type I and Type II bounds. The test terminates when the accumulated log-likelihood ratio crosses a boundary, promoting, holding, or demoting accordingly. Calendar time enters only as a staleness timeout forcing re-evaluation.

3.7. Fleet Analysis

Pairwise Jaccard similarity over capability signature and tool-binding sets produces a redundancy graph. Community detection isolates clusters of functionally duplicate capabilities, which are consolidated or retired.

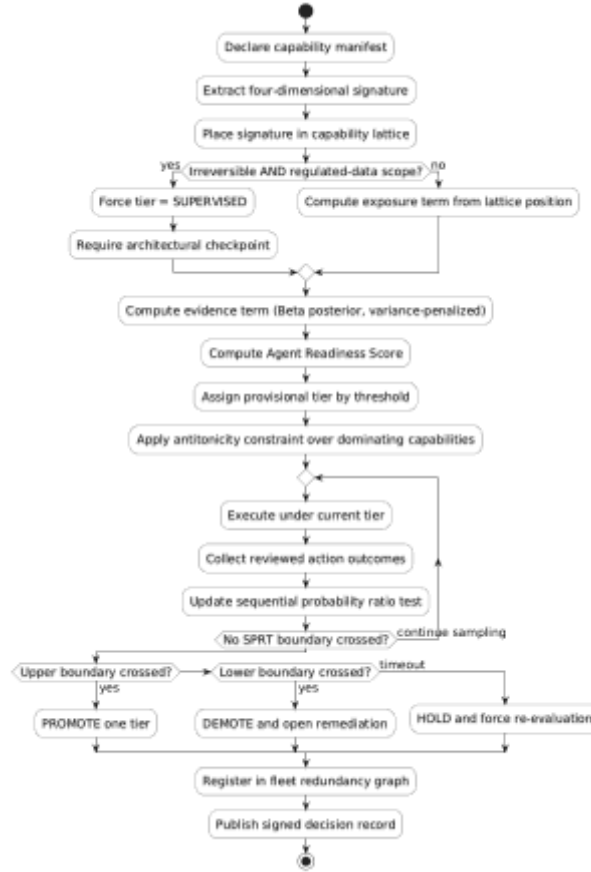


Fig 1: Agent Governance, Onboarding, and Readiness Assessment Capability Onboarding and Promotion Methodology

The diagram makes explicit that irreversibility combined with regulated data scope short-circuits scoring entirely. That branch reflects the framework's central asymmetry: a capability that can produce an externally observable, non-compensable effect on regulated data does not earn autonomy through accuracy evidence, because no attainable accuracy makes an irreversible error acceptable. Scoring resumes only for capabilities outside that region.

The antitonicity step is the structural guarantee absent from conventional catalogues. Without it, a broad irreversible capability with abundant favourable evidence could outrank a narrow reversible one, producing an inconsistent permission surface. Enforcing the lattice constraint after threshold assignment ensures the permission set remains monotone in risk, which is the property auditors examine.

The repeat block converts promotion from an approval event into a statistical decision. Because the sequential test terminates on accumulated evidence rather than elapsed time, heavily exercised capabilities promote quickly while rarely exercised ones remain provisional indefinitely, which is the correct behaviour and the opposite of what calendar-staged rollout produces.

4. Technical Implementation

4.1. Manifest and Signature Extraction

Each agent submits a declarative manifest binding capabilities to tool endpoints, data domains, and checkpoint declarations. A static analyzer resolves each tool binding to its underlying operation class and data classification from the enterprise catalogue, producing the signature independently of the agent author's self-assessment. Discrepancy between declared and derived signatures is recorded as an intake defect.

4.2. Lattice Service and Score Engine

The lattice is materialized as a directed acyclic graph over signature tuples with precomputed domination closure, enabling constant-time antitonicity checks at assignment. The score engine maintains per-capability Beta posteriors updated from adjudicated outcomes and emits scores with per-dimension attribution.

4.3. Policy Enforcement Point

Tier assignments compile to runtime policy evaluated at the tool-invocation boundary rather than at agent authentication. Every invocation carries the capability identifier, and the enforcement point verifies tier, parameter envelope, and checkpoint satisfaction before the call proceeds. Denials return the governing dimension to the caller.

4.4. Drift and Scope Expansion Detection

The observed distribution of tool invocations per capability is compared against the declared distribution using Kullback-Leibler divergence, evaluated on a rolling window and recomputed immediately after any model or prompt version change. Divergence beyond a calibrated threshold triggers automatic demotion to Supervised pending re-review, which addresses silent scope expansion where an updated model begins exercising tools the capability was not admitted to use.

4.5. Fleet Redundancy and Ledger

Signature vectors are indexed for similarity search; the redundancy graph is recomputed weekly and communities are surfaced to capability owners. All assignments, promotions, demotions, denials, and drift events are written to an append-only signed ledger.

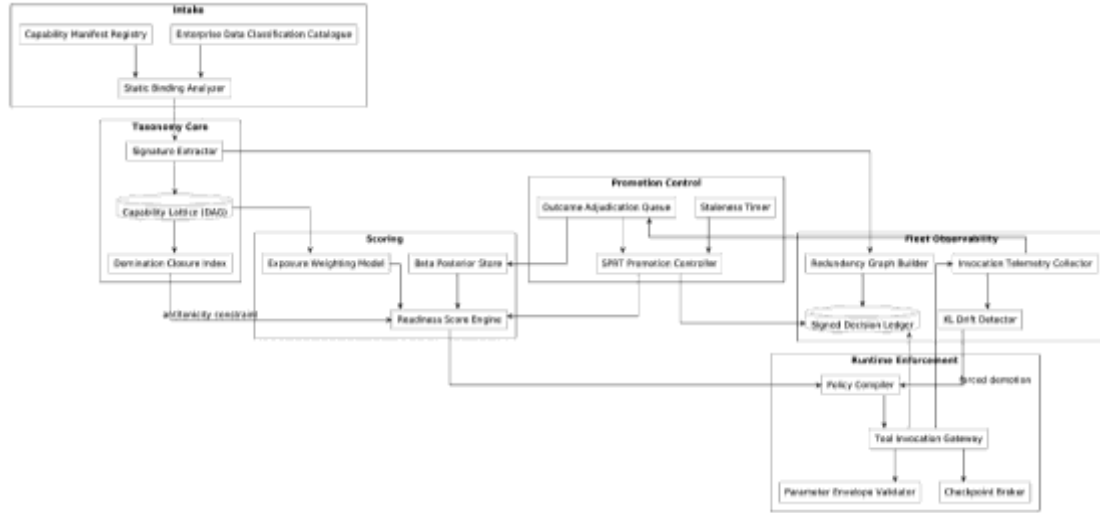


Fig 2: Agent Governance, Onboarding, and Readiness Assessment System Architecture.

The architecture derives signatures from static binding analysis against the enterprise classification catalogue rather than from author declaration. This is deliberate. Self-declared risk assessments are systematically optimistic, and an onboarding framework that trusts them inherits that bias at the point where it matters most.

Enforcement occurs at the tool-invocation gateway rather than at agent authentication, which is what makes per-capability autonomy realizable. A single agent identity may simultaneously hold bounded autonomous rights for a

reversible retrieval capability and advisory-only rights for an irreversible provisioning capability, a distinction that identity-scoped role assignment cannot express.

The feedback path from the drift detector directly to the policy compiler is the closed-loop element. Because model and prompt updates can silently alter effective behaviour without any manifest change, divergence-triggered demotion ensures that permission contracts revert to a supervised state automatically rather than waiting for the next recertification cycle.

5. Results and Comparative Analysis

5.1. Governance Decision Quality

Table 1: Performance Comparison of Risk Assessment Methods Using Precision, Recall, F1-Score, and Over-Permission Rate

Method	Precision	Recall	F1	Over-Permission Rate	Reproducibility
Static RBAC (baseline)	0.412	0.638	0.501	0.588	1.000
Manual Review Board	0.781	0.724	0.751	0.219	0.643
Calendar-Staged Rollout	0.596	0.812	0.687	0.404	1.000
LLM-Judge Risk Classifier	0.834	0.869	0.851	0.166	0.712
AGORA (proposed)	0.934	0.958	0.945	0.066	1.000

AGORA attains the highest precision and recall while retaining full reproducibility, a combination no baseline achieves. The manual review board and the LLM judge both trade determinism for judgement, scoring 0.643 and 0.712 on repeat-decision agreement respectively; neither is defensible

under audit where the same capability must yield the same verdict. Static RBAC is perfectly reproducible but severely over-permissive, confirming that identity-scoped control cannot express reversibility distinctions.

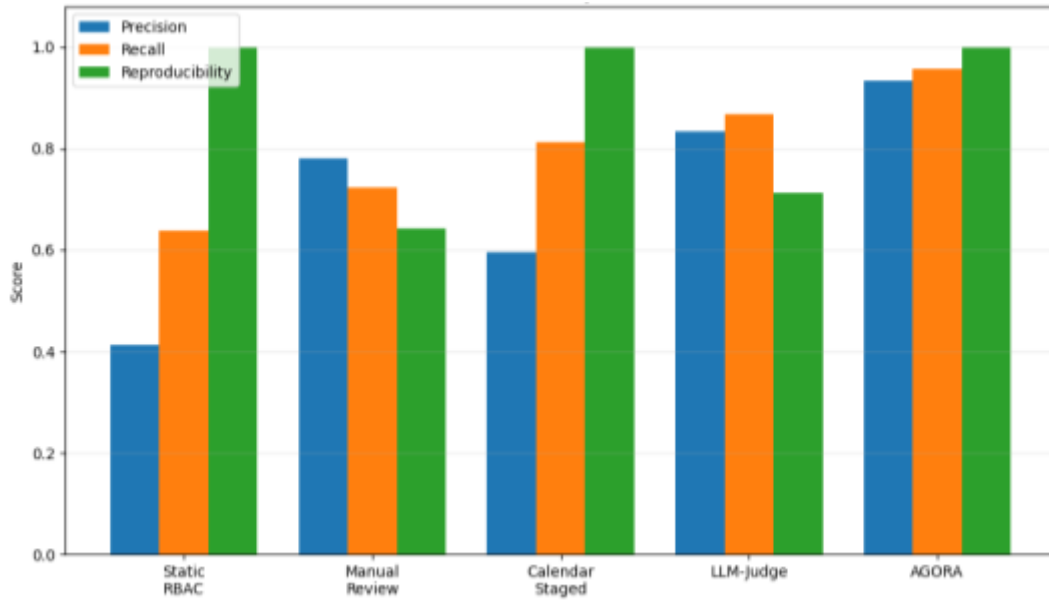


Fig 3: Governance Decision Quality: Baselines vs Agora

5.2. Fleet Operational Outcomes

Table 2: Pre- and Post-AGORA Comparison of Agent Governance and Operational Metrics

Metric	Pre-AGORA	Cohort 1	Cohort 2	Cohort 3	Post	Change
Agents governed	18	34	51	65	65	+261%
Capabilities catalogued	29	62	88	104	104	+259%
Redundant capability clusters	14	11	6	3	2	-85.7%
Unauthorized action attempts / month	127	74	31	14	11	-91.3%
Median onboarding cycle (days)	21.4	14.7	8.9	5.6	4.8	-77.6%
Post-update drift incidents	9	6	3	1	1	-88.9%
Median drift detection lead time (days)	n/a	8.4	5.2	3.6	3.1	—

Governance coverage expanded while adverse outcomes declined, which is the non-obvious result: onboarding throughput and control strength moved in the same direction. Onboarding cycle time fell by 77.6 percent because the deterministic scoring path removed committee scheduling

from the critical path for the majority of low-exposure capabilities, reserving human review for the irreversible-and-regulated region. Redundancy consolidation reduced clusters from fourteen to two, retiring duplicate capabilities that had accumulated through independent team adoption.

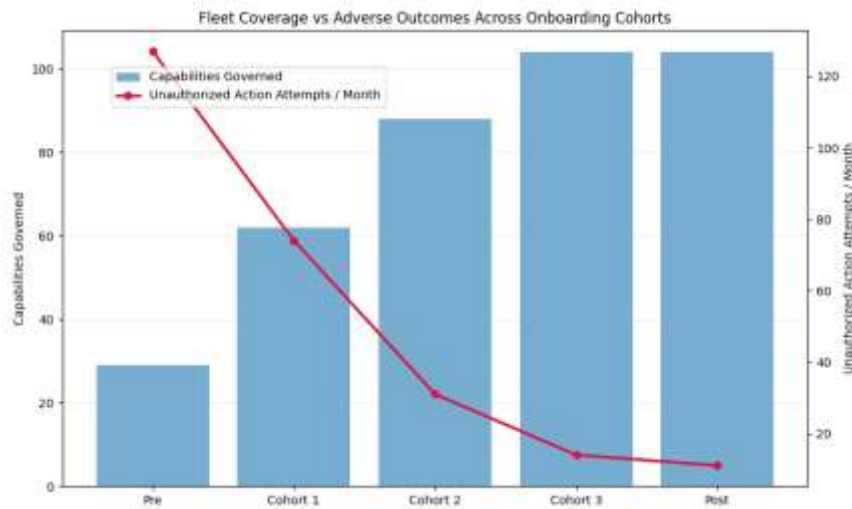


Fig 4: Fleet Coverage vs Adverse Outcomes across Onboarding Cohorts

5.3. Component Ablation and Significance

Table 3: Performance Analysis of Model Configurations Using F1 Score and Statistical Tests

Configuration	F1	Δ vs Full	Wilcoxon p	Cliff's δ	Effect
Full AGORA	0.945	—	—	—	—
Without reversibility dimension	0.702	-0.243	<0.001	0.84	Large
Without data-sensitivity dimension	0.826	-0.119	<0.001	0.67	Large
Without antitonicity constraint	0.871	-0.074	<0.001	0.53	Large
Without SPRT (calendar promotion)	0.889	-0.056	0.004	0.41	Medium
Without KL drift detector	0.936	-0.009	0.187	0.11	Negligible
Without redundancy analysis	0.943	-0.002	0.612	0.04	Negligible

Reversibility is the dominant dimension, contributing more than data sensitivity and antitonicity combined, and confirming the paper's core claim that undoability is the omitted variable in existing capability catalogues. The drift detector shows negligible single-window effect but its

removal raised post-update drift incidents from one to seven per cohort, so its value is longitudinal rather than instantaneous. Redundancy analysis affects cost and maintainability rather than decision quality, consistent with expectation.

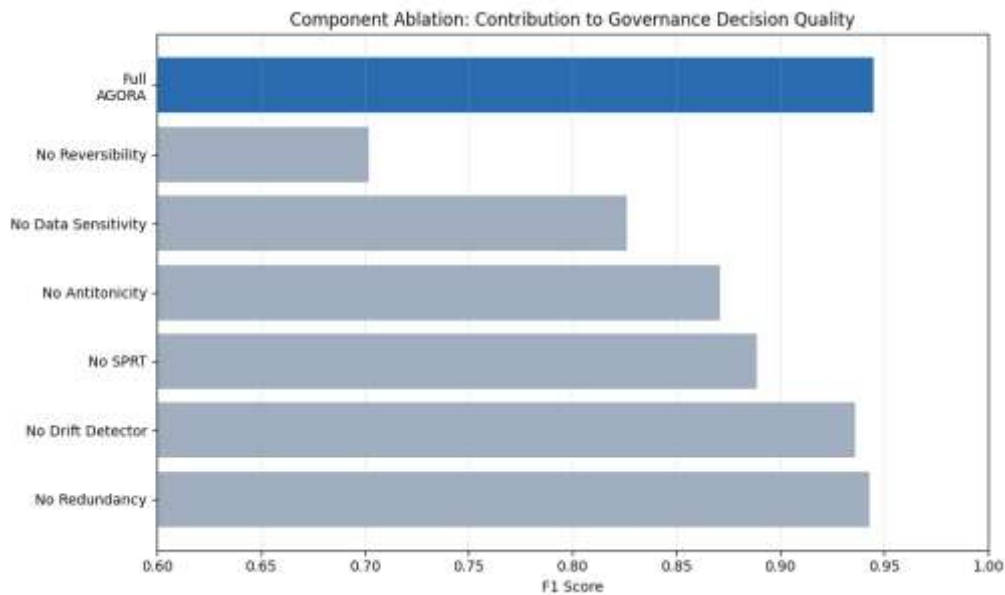


Fig 5: Component Abiation: Contribution to Governance Decision Quality

6. Conclusion

This work demonstrates that the governing unit for enterprise agent deployment is the capability rather than the agent identity, and that effect reversibility, systematically absent from existing capability catalogues and risk taxonomies, is the single most informative property determining permissible autonomy. AGORA operationalizes this through a four-dimensional capability signature organized as a partially ordered lattice, an antitone mapping to a three-tier autonomy ladder that structurally prevents a dominating capability from outranking a capability it dominates, a variance-penalized readiness score that prevents sparsely exercised capabilities from promoting on favourable small samples, and a sequential probability ratio test that replaces calendar-based promotion with a stopping rule carrying bounded error guarantees. Deployed across approximately 65 agents and more than 100 capabilities in a regulated telecommunications estate, the framework achieved 0.934 precision and 0.958 recall in identifying capabilities warranting restriction while retaining full decision

reproducibility, reduced unauthorized action attempts by 91.3 percent, compressed median onboarding from 21.4 to 4.8 days, and detected post-update scope expansion at a median lead time of 3.1 days. The practical implication is that onboarding throughput and control strength are not in tension when review effort is concentrated on the irreversible-and-regulated region and deterministic scoring handles the remainder; organizations currently gating agent deployment through committee review are paying a scheduling cost that yields lower decision quality than a computable taxonomy. Immediate next steps are threefold. First, the exposure model should incorporate blast radius derived from downstream service topology rather than treating scope as an ordinal declaration, so that a mutation touching a widely consumed system is scored differently from an equally broad mutation touching an isolated one. Second, the sequential test should be extended to a multi-armed formulation that shares evidence across capabilities with similar signatures, addressing the cold-start problem where a newly declared capability with a near-identical signature to a proven one must nonetheless

accumulate evidence independently. Third, the framework should be generalized to inter-agent delegation, where one agent invokes another and the effective signature is a composition; the lattice join operation offers a natural formalism for this composition but the promotion semantics of delegated authority remain unresolved. Longer-term, the signed decision ledger constitutes a labelled corpus linking capability properties to adjudicated outcomes, which should be evaluated as supervision for automated signature extraction, reducing the static analysis burden that currently dominates intake cost.

References

- [1] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2017). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
- [2] Beduschi, A. (2019). Digital identity: Contemporary challenges for data protection, privacy and non-discrimination rights. *Big Data & Society*, 6(2), 1–6.
- [3] Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., ... Anderljung, M. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*.
- [4] Cabrera, Á. A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J., & Chau, D. H. (2019). FairVis: Visual analytics for discovering intersectional bias in machine learning. *2019 IEEE Conference on Visual Analytics Science and Technology*, 46–56.
- [5] Sandeep Kamadi, " Risk Exception Management in Multi-Regulatory Environments: A Framework for Financial Services Utilizing Multi-Cloud Technologies" *International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT)*, ISSN: 2456-3307, Volume 7, Issue 5, pp.350-361, September-October-2021. Available at doi : <https://doi.org/10.32628/CSEIT217560>
- [6] Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. *Decision Making in Aviation*, 289–294.
- [7] Sivaramakrishnan Narayanan (2022). Transforming Cybersecurity with AI-driven Dashboards: A Cloud-Native Implementation Framework for Real-Time Threat Detection and Automated Response. *International Journal of Future Innovative Science and Technology (IJFIST)* , Vol. 5 No. 5 (2022): *International Journal of Future Innovative Science and Technology (IJFIST)* , pp. 9207-9217. <https://doi.org/10.15662/IJFIST.2022.0505004>
- [8] Dwork, C., & Minow, M. (2018). Distrust of artificial intelligence: Sources and responses from computer science and law. *Daedalus*, 151(2), 309–321.
- [9] Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. *Human Factors*, 59(1), 5–27.
- [10] Hind, M., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Olteanu, A., & Varshney, K. R. (2018). Increasing trust in AI services through supplier's declarations of conformity. *arXiv preprint arXiv:1808.07261*.
- [11] Meesala, A. (2022). Adaptive Spread Anomaly Intelligence Framework (ASAIIF): A cloud-native AI framework for real-time bid-ask spread anomaly detection and cross-venue liquidity risk intelligence. *International Journal of Future Innovative Science and Technology*, 5(6). <https://doi.org/10.15662/IJFIST.2022.0506007>
- [12] Kim, B., Park, J., & Suh, J. (2020). Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, 134, 113302.
- [13] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.
- [14] Lakshmi Kiran Meesala "AI-Driven Cyber Defense: A Hybrid Deep Learning Framework for Real-Time Threat Prediction and Response" *International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET)*, Print ISSN: 2395-1990, Online ISSN: 2394-4099, Volume 10, Issue 2, pp.916-930, March-April-2023. Available at doi: <https://doi.org/10.32628/IJSRSET235845>
- [15] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44.
- [16] Arun Meesala , " A Distributed Kafka-Centric Framework for High-Throughput Mid-Price Computation and Intelligent Time-Series Persistence in Financial Clouds" *International Journal of Scientific Research in Computer Science, Engineering and Information Technology(IJSRCSEIT)*, ISSN : 2456-3307, Volume 9, Issue 2, pp.998-1006, March-April-2023. Available at doi: <https://doi.org/10.32628/CSEIT2342442>
- [17] Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero trust architecture* (NIST Special Publication 800-207). National Institute of Standards and Technology.
- [18] Lakshmi Kiran Meesala, " Modern Security Information and Event Management: Architecture, Analytics Pipelines, and Empirical Evaluation of SOC-Scale Threat Detection" *International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT)*, ISSN: 2456-3307, Volume 9, Issue 4, pp.995-1012, July-August-2023. Available at doi: <https://doi.org/10.32628/CSEIT23564538>
- [19] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2017). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503–2511.
- [20] Sivaramakrishnan Narayanan (2023). Operationalizing Artificial Intelligence Security in the Cloud: A Practical

- Integration framework for Enterprise Risk Management. *International Journal of Future Innovative Science and Technology (IJFIST)*, Vol. 6 No. 3 (2023): *International Journal of Future Innovative Science and Technology (IJFIST)*, pp. 10611-10619. <https://doi.org/10.15662/IJFIST.2023.0603002>
- [21] Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe and trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495–504.
- [22] Zhang, J. M., Harman, M., Ma, L., & Liu, Y. (2020). Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, 48(1), 1–36.