KUBECON + CLOUDNATIVECON NORTH AMERICA 2026 · AI INFERENCE + AGENTIC

From Black-Box to Auditable AI-Grid

A deterministic governance layer with hybrid cloud

Andrew C. Hendrickson

Chief, Office of Engineering & Technology, FCC

Fatih E. Nar

Distinguished Chief Architect, Red Hat

THE SHIFT

AI is moving into the operating core of the network

Control loops
Closed-loop actions on live radio and core state

Spectrum allocation
Who gets which resources, and when

Slicing with QoS
Slices that carry service commitments

Customer issues
Automated triage and resolution

These networks carry 911 and priority traffic. Decisions made on them may need to be **explained to an investigator after the fact.**

AFTER AN INCIDENT

The questions an investigator asks

1. Which model, and which version, shaped the decision?
2. Which data and evidence informed it?
3. Was it operating inside its validated bounds?
4. Could a human override or halt it?

A model recommendation leaves the operator's responsibility in place. Opaque, unreproducible outputs make all four questions hard to answer.

WHAT "DETERMINISTIC" MEANS HERE

Bounded reproducibility

HOLD FIXED

The input

The rule-base version

The corpus snapshot



REPRODUCE

The same rules invoked

Materially equivalent evidence

The same citations

Identical wording from the language model is out of scope. The citation trail is the audit artifact.

THE POLICY VIEW

Existing obligations already apply

In force today

- Reliability and resilience
- Emergency service continuity
- Lawful operation and privacy

These apply to an AI-driven outcome the same way they apply to a manual change.

Specifics follow evidence

- EU AI Act risk tiers
- Financial model risk management
- ISO/IEC 42001 AI management

Each arrived after enough operating experience to justify targeted requirements.

Starting point: **traceable decisions, structured logging, risk-tiered approvals.**

TWO JOBS, ONE PLATFORM

AI for Networks and Networks for AI

Dimension	AI for Networks	Networks for AI
Objective	Operational efficiency and performance	External services, platform revenue
Core data	Operator telemetry	Customer workloads
Dependencies	Observability, MLOps, operations	Tenancy, SLAs, platform APIs
Main risk	Model trust, data quality	Demand, utilization
AI-RAN Alliance	AI-for-RAN	AI-on-RAN, AI-and-RAN

One hybrid-cloud platform can carry both. **The governance layer is what they share.**

ARCHITECTURE

Where governance sits in AI-RAN

Layer 4

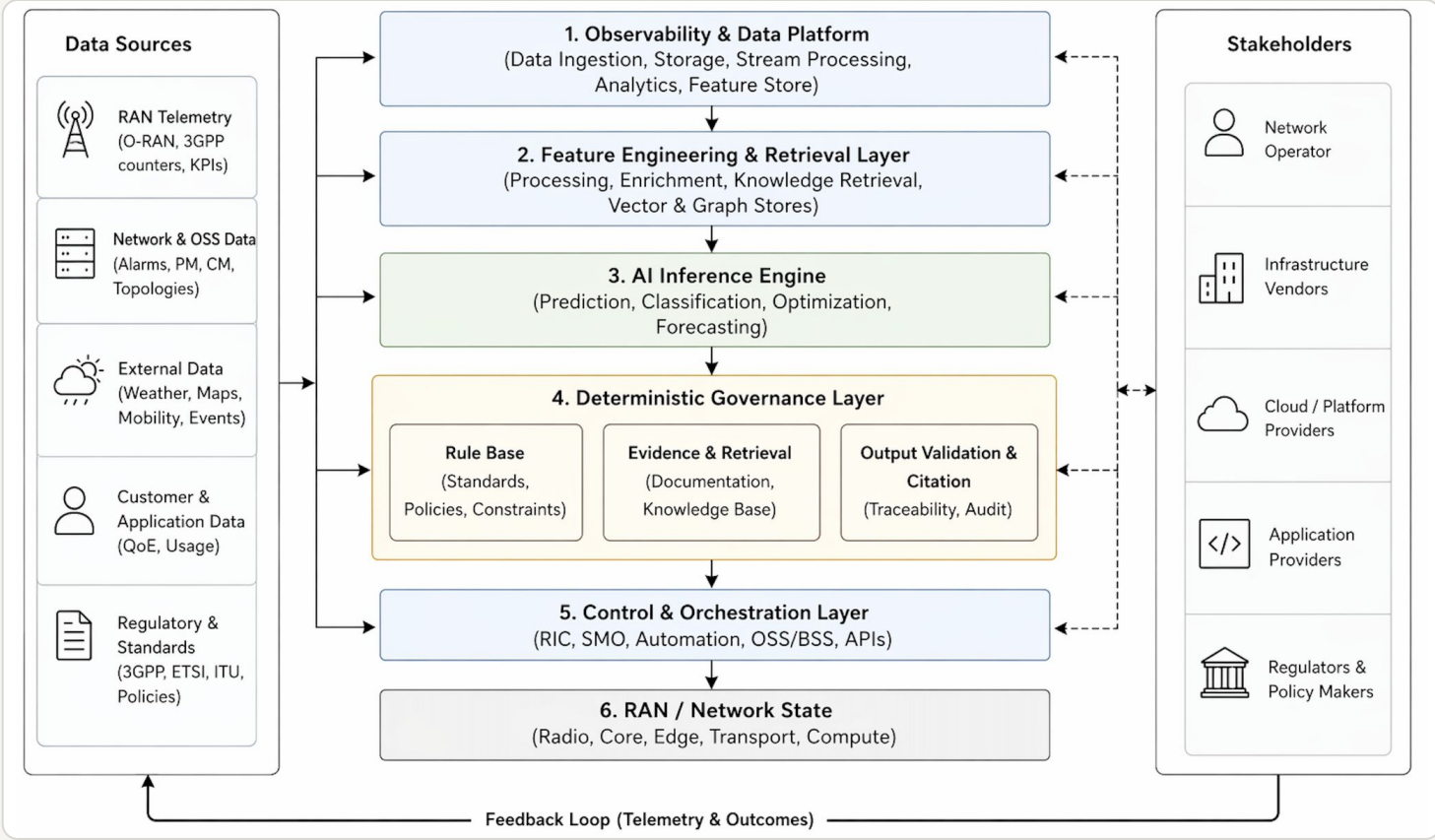
Deterministic governance sits between inference and control

Three parts

Rule base, evidence retrieval, output validation with citation

Closed loop

Outcomes flow back as telemetry



THE CONTRAST

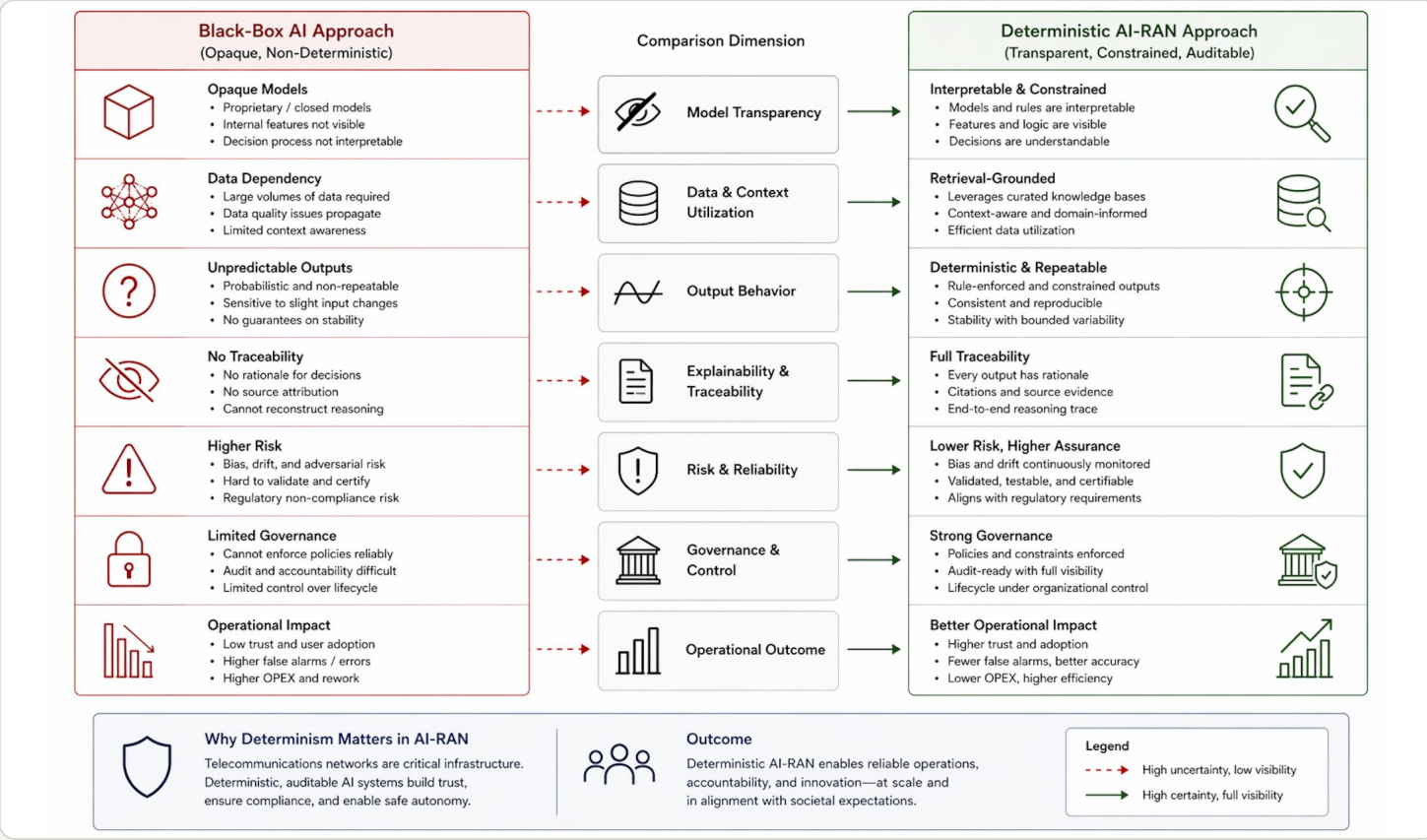
Black-box AI and the deterministic pattern

Black box

Data in, action out. The reasoning stays inside the model.

Deterministic

Rules and retrieved evidence bound the synthesis. Every output carries a citation trail.



THE PATTERN

A five-stage deterministic pipeline

1 Intent

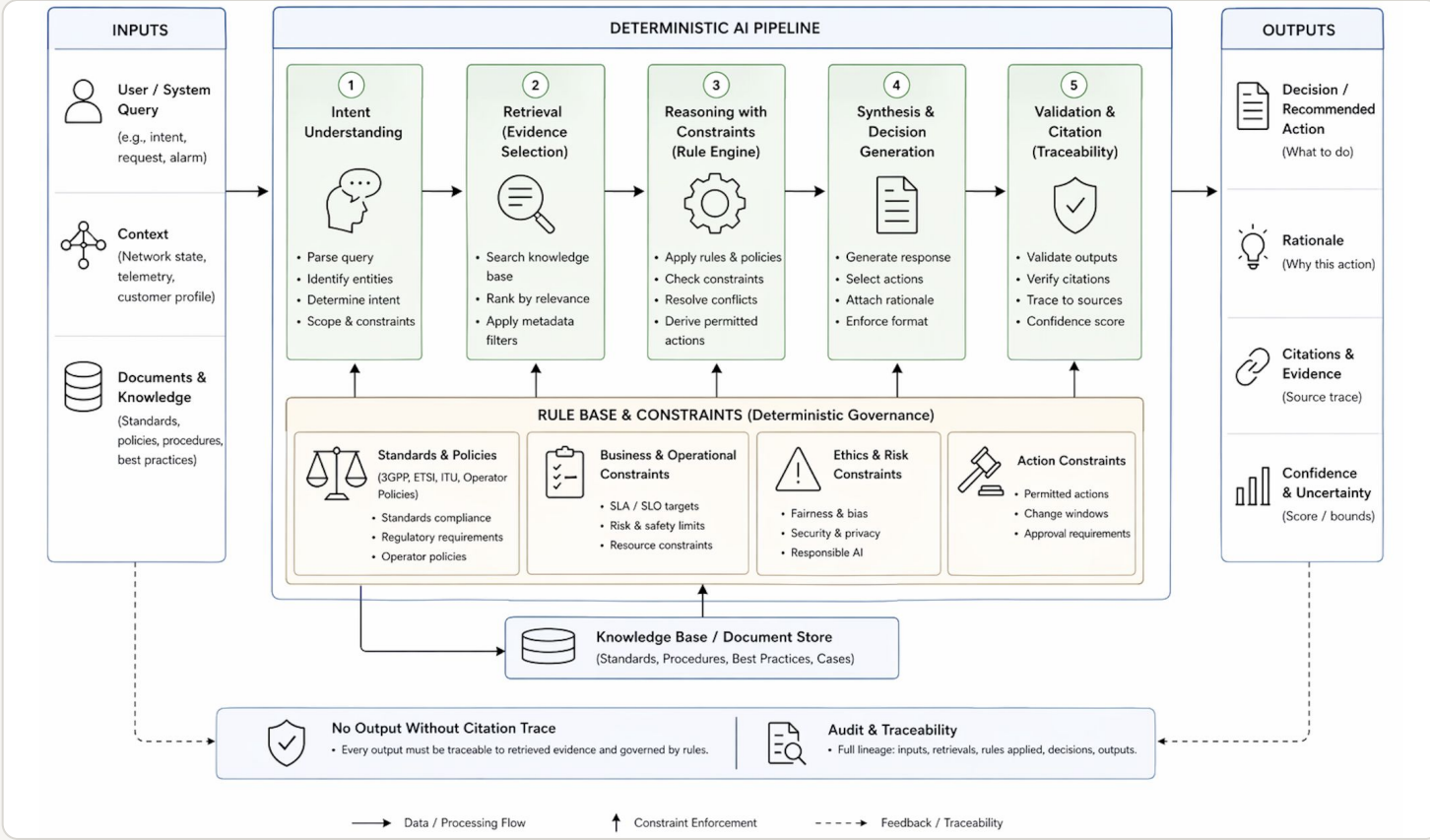
2 Evidence retrieval

3 Rule-bounded reasoning

4 Synthesis and decision

5 Citation-validated output

No output without a citation trace.



WALKTHROUGH

Putting a cell sector to sleep

MODEL PROPOSES Switch off one sector in a forecast low-traffic window

OUTCOME A · PERMIT

- No emergency sessions on the sector
- Neighbors have spare capacity
- Handover failures under threshold
- Inside an approved change window

Instruction issued with citation trail and a scheduled reactivation

OUTCOME B · ESCALATE

- An active emergency session, or
- Neighbor handover failures above threshold

No instruction. Escalation cites the blocking rule, the evidence and the rule-base version, then goes to human review.

Both paths leave the same audit artifact: **the citation trail.**

DEPLOYMENT MOTTOS

Defined behavior when inputs fall short

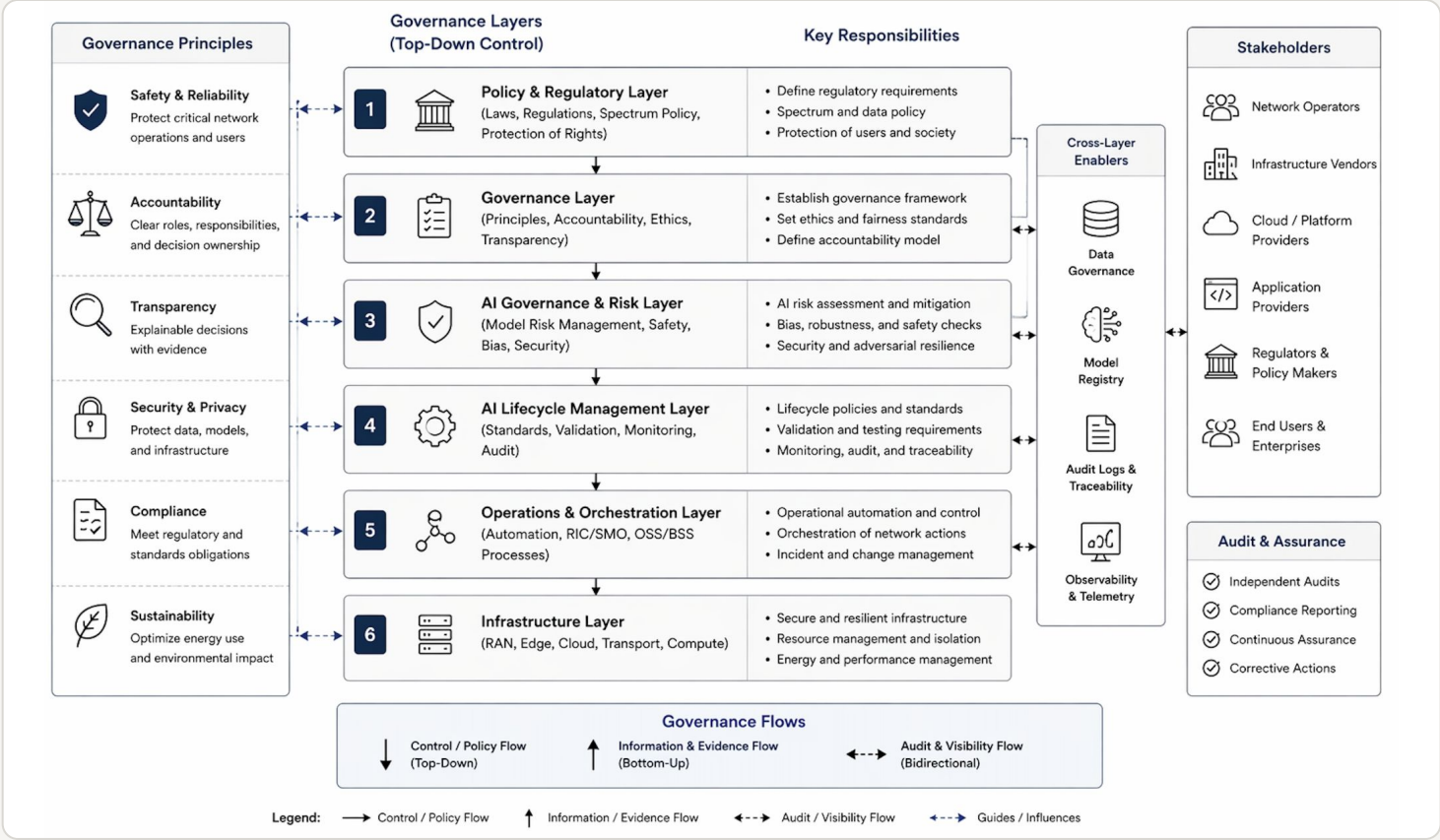
Condition	Defined behavior
No relevant evidence retrieved	Return “insufficient evidence” and escalate; do not synthesize
Applicable rules conflict	Apply documented precedence; log conflict and resolution
Output cites unsourced material	Reject before delivery; log a synthesis failure
Rule or corpus version missing	Refuse to answer without provenance
Latency budget exceeded	Validated partial answer or escalate; never strip citations

Behind the table: a versioned rule base with explicit precedence, a content-hashed evidence corpus, and claim-level validation before return.

GOVERNANCE MODEL

Six governance layers for AI-RAN

1. Policy and regulatory
2. Governance
3. AI governance and risk
4. AI lifecycle management
5. Operations and orchestration
6. Infrastructure



REFERENCE PATTERN

Layers mapped to cloud-native projects

Layer	What it does at runtime	Candidate projects
1 Policy, regulatory	Obligations kept as versioned rules-as-code	Git, Argo CD or Flux
2 Governance	Block deploys missing owner or hashes	OPA Gatekeeper or Kyverno
3 AI risk	Evaluate rules per decision, keep decision logs	OPA (Rego)
4 AI lifecycle	Promote model, rules and corpus together	GitOps: Argo CD or Flux
5 Operations	Decision traces carry rule and citation IDs	OpenTelemetry
6 Infrastructure	Energy per pod, AI overhead counted	Kepler

NOTE: The project names given here are candidates, and equivalent projects fit the same slots.

LIFECYCLE

GitOps for model, rules and corpus

rules/

Rego policies, precedence order, semantic version

corpus/

Snapshot manifest of content hashes

models/

Image digest, model card, owner

DECISION RECORD (ILLUSTRATIVE)

```
{
  "decision": "d-0412",
  "rules": "rules@v3.4.1",
  "corpus": "sha256:9f2c…e1",
  "model": "energy-opt@sha256:41ab…",
  "evidence": ["ev-118", "ev-207"],
  "outcome": "escalate",
  "blocked_by": "R-EMERG-001"
}
```

Replay: check out the three pinned versions, rerun, and compare against the equivalence class.

AGENTIC AI

Agents move the trust boundary

The decision moves

Decisions now come from a runtime mix of model, tools and plan

Constrain the action space

Agents call only rule-sanctioned operations, each with a cited justification

Keep the other controls

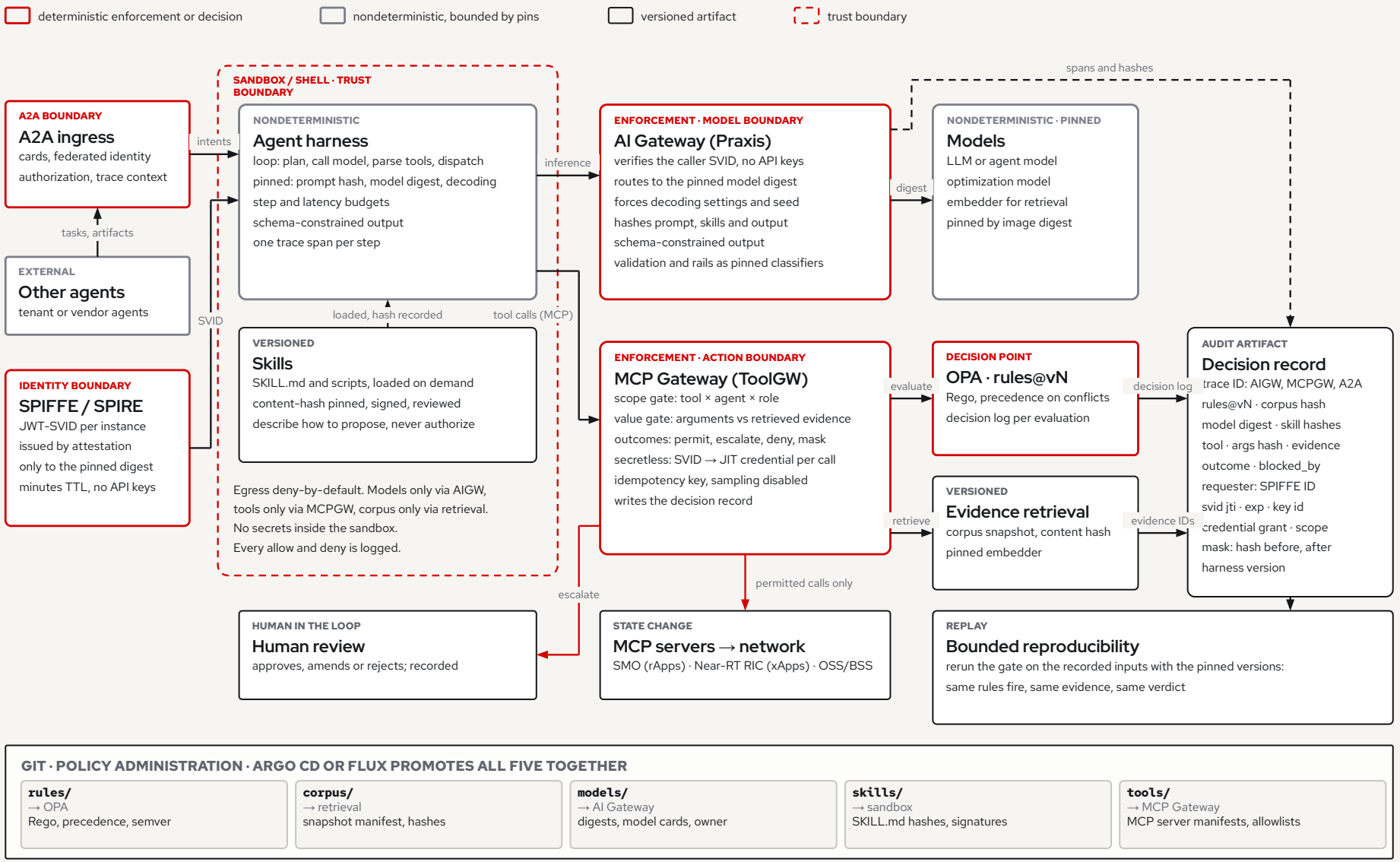
Tool allowlists, input sanitization, retrieval isolation, human override, adversarial testing

EXAMPLE FROM PRACTICE (O-RAN WG)

MX-AI agents in the O-RAN SMO: 100% decision-action accuracy, 8.8 s end to end, over 50 queries. A strong demonstration. Production evidence is still ahead.

AGENTIC AI

Agentic Blueprint for Determinism



MATURITY PLAN

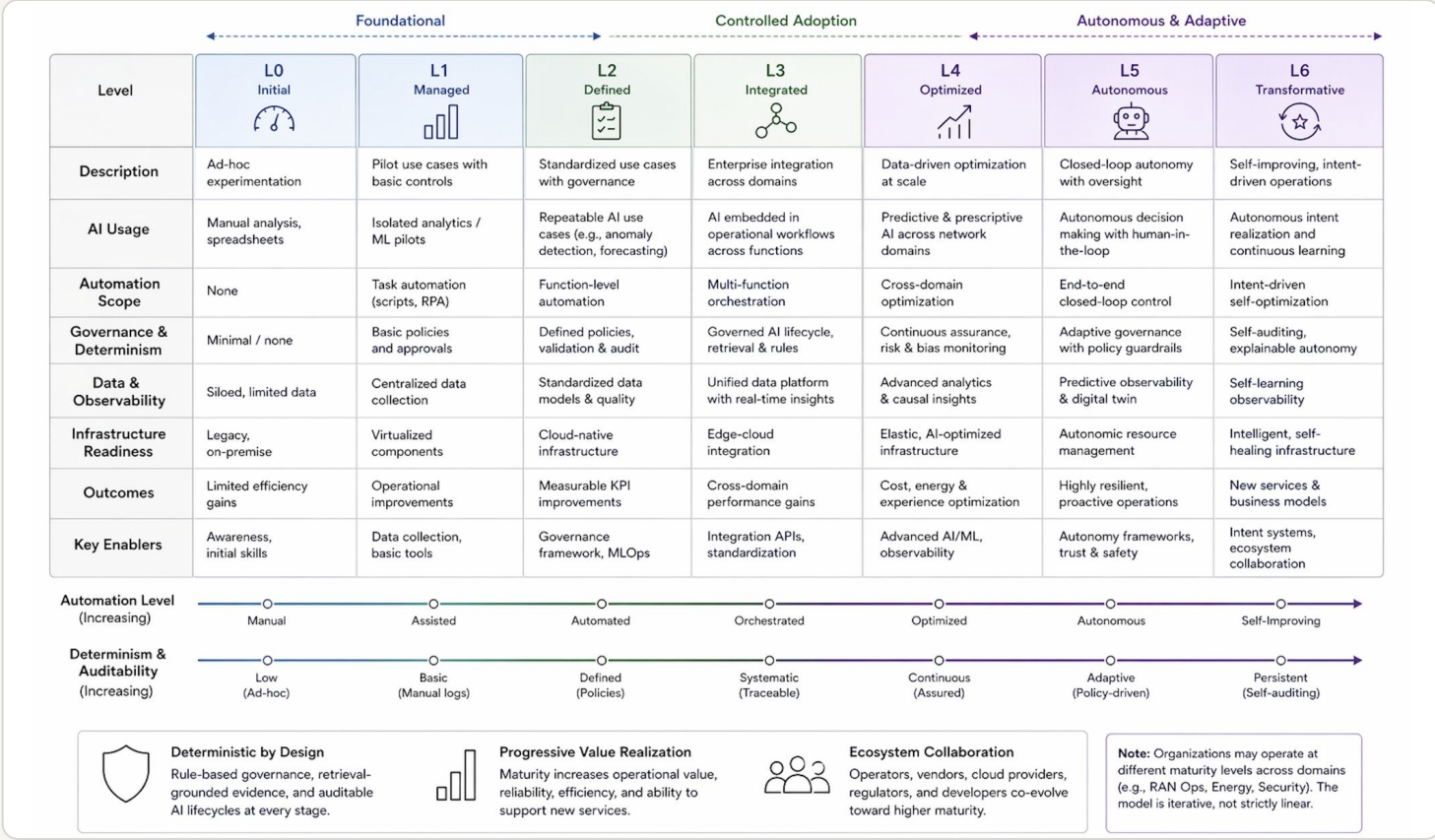
A progressive path : L0 -> L6*

L0 to L5
TM Forum Autonomous Networks levels, unchanged

L6 (author-proposed)
L5 autonomy plus cited governance for every regulated decision

Hosted-Workload Profile
Tenant isolation, verifiable SLAs, tenant-aware audit; L3 and above

L6* = L6 + HWP
Also author-proposed. TM Forum defines no L6.



PROMISED REWARD

Key questions for any energy claim

Ask	Expect to see
What is included?	Compute, storage, network, cooling
Against what baseline?	A clear non-AI or prior state
Traffic normalized?	Energy per bit and total energy
Service quality held?	QoS, SLA and outage metrics
Reproducible?	Over time, across scenarios

Kepler and OpenTelemetry put the AI's own power draw in the same report as the savings.

~945 TWh

Projected data-center electricity in 2030, up from about 415 TWh in 2024 (IEA)

> 2/3

Share of the mobile industry's footprint that is Scope 3 (GSMA)

THE ADOPTION GAP

Four barriers, and where to start

Economic
Governance ROI is defensive and harder to fund

Organizational
Network, AI, compliance and legal each own a piece

Governance
Audit specifics for AI decisions are still forming

Tooling
No integrated telco-grade governance stack yet

START AT L2 & L3 Versioned rules, structured decision logs, risk-tiered approvals. Walk towards L6* with Determinism.

Questions

Paper: *AI-RAN with Deterministic Patterns & Practices*, Hendrickson and Nar

Andrew C. Hendrickson

Chief, Office of Engineering & Technology, FCC

Fatih E. Nar

Distinguished Chief Architect, Red Hat