Technology Stewardship

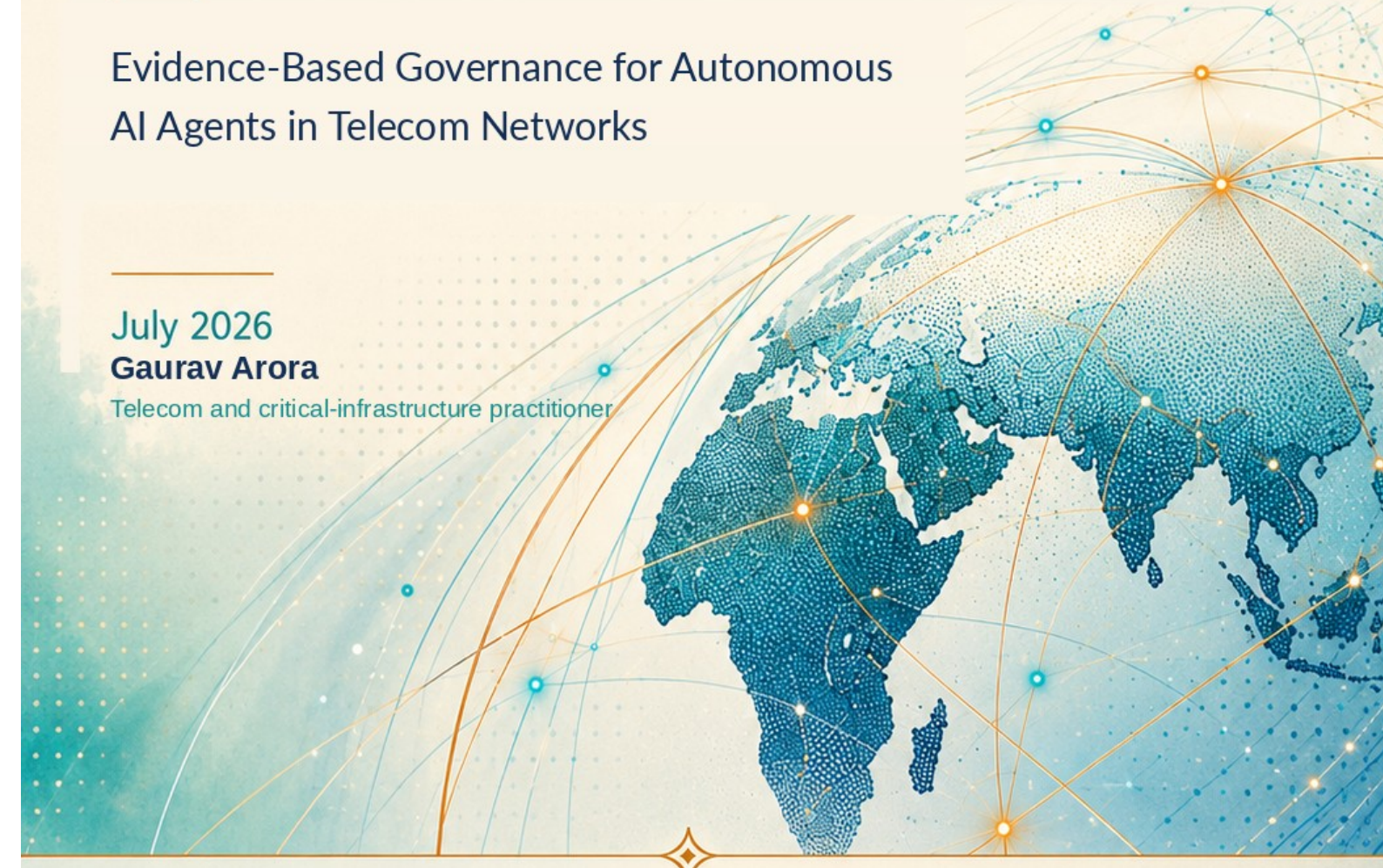
INSTITUTE • GOVERNING AUTONOMOUS TECHNOLOGY

INSTITUTE FOR TECHNOLOGY STEWARDSHIP

ISSUES PAPER No. 2

Beyond the Sandbox

Evidence-Based Governance for Autonomous AI Agents in Telecom Networks



July 2026

Gaurav Arora

Telecom and critical-infrastructure practitioner

Ensuring technology remains in service of humanity — not the other way around.

Executive Summary

Author: Gaurav Arora | Institute for Technology Stewardship | Research cut-off: 18 July 2026

Keywords: telecommunications; autonomous agents; AI governance; Continuous Shadow Deployment; human oversight; critical infrastructure; Global South

Telecommunications operators are moving from AI-assisted analysis toward systems that can diagnose faults, select tools and, within bounded domains, execute changes without per-action approval. Deutsche Telekom reports that its production RAN Guardian Agent autonomously triggered more than 100 remediation actions in its first month; the broader MINDR cross-domain architecture was announced for later 2026, although public material does not disclose the approval workflow for each action. GSMA and TM Forum programmes likewise treat agentic AI and Level 4 autonomous networks as practical industry objectives.[1, 2, 3, 4]

This changes the governance question. A predictive or advisory model can be reviewed before a person acts. An agent with delegated authority can change a live network at machine speed, inherit permissions and interact with other agents and tools. The control problem is therefore not only whether the system performs well, but whether authority has been earned, bounded, observed and made safely reversible.

Sandboxes, simulations and predictive digital twins remain indispensable: they test capability, robustness and candidate actions under defined conditions. Yet no controlled environment reproduces every live dependency, legacy interaction, traffic pattern, permission boundary or multi-agent conflict. Pre-deployment validation is necessary, but not sufficient evidence for autonomous authority.[5, 6]

Continuous Shadow Deployment (CSD) is proposed as the missing evidentiary layer. Conventional shadow testing evaluates an output that is not served; CSD evaluates an agent's claim to use tools, permissions and operational authority. Because that claim is stateful, the evidence needed depends on the realism of the feedback environment. CSD therefore combines observational shadowing, twin- or replay-mediated feedback and bounded live canary execution. Authority is granted progressively by action class and network domain, supported by a vendor-neutral decision-event record, risk-based adjudication, independent runtime controls and automatic contraction when evidence deteriorates.

CSD is not a legal requirement or a replacement for sandboxes, twins, human oversight or post-market monitoring; it connects them into a production assurance lifecycle. It extends ITS Issues Paper No. 1: the earlier paper asks whether an operator controls the data, model, jurisdiction and decision record; this paper asks whether delegated authority is reconstructable, earned against evidence and interruptible in time. Together, the six questions address the data-governance and authority-governance layers of telecom AI.

The stakes are concrete. In September 2025, a planned firewall-upgrade change on the Optus network disrupted Triple Zero emergency calling for more than 14 hours. Of 605 unique callers, only 150 connected; an independent review related two fatalities to the outage. This was not an AI incident. It was a human-governed production change failure - and it shows why machine-speed authority requires a higher evidentiary threshold.[21, 22]

THREE GOVERNANCE-READINESS QUESTIONS

1. Can the operator reconstruct every consequential decision and action?
2. Has authority been earned against predefined, independently adjudicated thresholds?
3. Can a qualified human intervene within the available intervention window and return the system to a safe state?

If an operator cannot reconstruct the event, demonstrate how authority was earned, and intervene within the available window, it is not ready to delegate consequential authority.

Methodology, Scope and Limitations

This paper is a desk-based policy and operational analysis prepared in July 2026. It draws on publicly available operator announcements, vendor and industry material, standards initiatives, government guidance, primary legal texts, an independent telecom-incident review and a regulator statement. Public announcements are treated as evidence of direction and claimed capability, not as independent proof of maturity, safety or governance effectiveness.

The analysis distinguishes offline tests, regulatory sandboxes, predictive twins, production shadow operation and authorised execution. They may coexist in one lifecycle but are not substitutes. Company examples show industry direction, not uniform architecture, autonomy or maturity.

Continuous Shadow Deployment is an ITS proposal. Shadow testing, canary releases and parallel runs are established practices; the contribution is to change the object of evaluation from an unserved prediction to a claim to use tools, permissions and operational authority. Because an authority claim is stateful, its evidentiary validity depends on the realism of feedback, available permissions and the consequences of preceding actions. CSD is therefore not simply a sequence of release techniques: it is an architecture that matches delegated authority to the strength and realism of the evidence supporting it.

This paper does not provide legal advice. EU AI Act analysis is system-specific: Annex III point 2 concerns AI used as a safety component in the management and operation of critical digital infrastructure, while other telecom systems require separate classification. Regulatory obligations, application dates and organisational roles should be verified for each deployment and jurisdiction.[7, 8, 9, 10]

1. From Automation to Delegated Authority

Telecom networks have used automation for decades. Rules engines, policy controllers, self-organising network functions and closed-loop assurance already adjust parameters or trigger predefined workflows. Agentic AI changes the control surface because the system may interpret an objective, select tools, plan a sequence of actions, interact with other systems and adapt its next action based on intermediate results.

The key distinction is delegated authority. An agent may be permitted to read alarms, query topology, correlate events, generate a configuration, open a ticket, execute a reversible parameter change or isolate a network element. Each additional permission changes the potential blast radius. Governance must therefore attach not only to the model, but to the complete agentic system: identity, tools, data, memory, permissions, orchestration, policies, human interfaces and execution environment.[11, 12]

Operating mode	Typical system behaviour	Human role	Primary governance concern
Assistive	Retrieves information or summarises alarms	Human interprets and acts	Accuracy, data protection and over-reliance
Advisory	Produces diagnosis or recommended action	Human approves or rejects each action	Explanation, verification and decision accountability
Supervised execution	Executes selected actions after approval	Human authorises each action or batch	Permission scope, safe execution and rollback
Bounded autonomy	Executes approved classes of action within limits	Human monitors and handles exceptions	Runtime policy, intervention window and drift
Cross-domain autonomy	Coordinates actions across agents and network domains	Human oversees objectives and emergency control	Emergent behaviour, cascading effects and accountability opacity

Industry signals indicate that this progression is underway. Deutsche Telekom describes MINDR as a multi-agentic system operating across radio, transport and core domains, and reports that its RAN Guardian Agent triggered more than 100 autonomous remediation actions during its first month in live operation. GSMA has described agentic AI as capable of detecting service degradation, rerouting traffic and notifying users, while TM Forum programmes now address governed, zero-trust agent runtimes and Level 4 network operations.[1, 2, 3, 4]

These developments demonstrate operational possibility, not uniform maturity. Public announcements generally do not disclose the complete permission model, event-log schema, human intervention design, incident thresholds or independent assurance process. The governance task is therefore to define what evidence an operator should possess before delegated authority expands.

2. What Controlled Testing Proves - and What It Cannot

Controlled testing is indispensable. A sandbox can isolate a system from production, reproduce known fault patterns and verify that tools, prompts, policies and integrations behave as intended. Simulation can generate rare events. A digital twin can represent network state and, when sufficiently predictive, estimate how proposed actions may change that state. These capabilities reduce risk and should remain mandatory components of any serious deployment process.

The limitation is structural: a controlled environment validates behaviour against the conditions, dependencies and threat models represented within it. Live telecom networks contain unmodelled states, undocumented legacy behaviour, asynchronous updates, partial telemetry, human workarounds, vendor-specific semantics and traffic shocks. A system may therefore pass every laboratory test and still fail when the production context differs from the test model.

> **A governance principle**
>
> Testing demonstrates capability under specified conditions. Authority requires evidence that the system remains observable, bounded and reversible when those conditions are incomplete, changing or wrong.

3. Production Failure Modes That Sandboxes Commonly Underrepresent

3.1 Distribution and context shift

Traffic patterns, subscriber density, equipment state, weather, public events, attacks and operational priorities change continuously. Even where individual variables are represented in testing, their combinations may not be. The relevant question is not whether the agent has seen a similar event, but whether it recognises uncertainty and escalates rather than confidently optimising against a false model of the situation.

3.2 Permission inheritance and tool expansion

An agent frequently receives access through service accounts, orchestration platforms or tools originally designed for human engineers. Permissions may be broader than the agent’s intended function. A model update, new tool or changed workflow can therefore expand effective authority without a formal governance decision. Agent identity, short-lived credentials, least privilege and tool-level policy enforcement are central controls, not implementation details.[11]

3.3 Multi-agent interference

One agent may optimise radio capacity while another reduces energy consumption and a third protects a service-level objective. Each action can be locally rational while the combination is unstable. Coordination failures may also arise when agents use different state representations, refresh intervals or confidence thresholds. End-to-end observability must therefore cover agent-to-agent and agent-to-tool interactions rather than only final outputs.[3, 11, 13]

3.4 Goal conflict and metric gaming

A narrow objective such as reducing mean time to repair can encourage aggressive remediation that increases change risk. An uptime objective can conflict with cybersecurity isolation, energy efficiency or equitable service distribution. Operational goals must be expressed with constraints, prohibited actions and priority rules, and the governance system must record when competing objectives were resolved.

3.5 Temporal compression

Machine-speed action can collapse the time available for human oversight. A human may formally possess override authority while lacking enough time, information or interface clarity to use it. Oversight must therefore be measured through intervention windows, alert-to-comprehension time, action latency and safe-stop performance, not merely through the presence of a named overseer.

3.6 Feedback-loop and recovery risk

Once an action changes the network, the resulting telemetry becomes new input. The agent may interpret the consequences of its own action as evidence for additional changes, creating a reinforcing loop. Safe recovery requires state checkpoints, idempotent actions where possible, rollback plans, rate limits and conditions that automatically return the agent to recommendation or shadow mode.

3.7 Operational precedent: when established change governance still fails

On 18 September 2025, a planned network change associated with a firewall-upgrade programme caused Triple Zero emergency calls to fail across South Australia, Western Australia, the Northern Territory and parts of far west New South Wales. Service was not restored until 14:34, more than 14 hours after the change began. Of 605 unique callers who attempted to reach emergency services, only 150 connected, leaving 455 unconnected. The independent review related two fatalities to the outage.[21, 22]

The review identified at least ten mistakes across planning, change instructions, method selection, peer review, risk classification, implementation, alert handling and monitoring. It concluded that processes and controls existed, but that the correct process was not followed and the controls were not applied properly.[21]

This was not an AI incident. It occurred within an established, human-governed environment involving formal change requests, peer review, risk classification, network monitoring and vendor participation. Its relevance is not that autonomous agents caused the failure, but that severe public consequences can arise even where conventional controls already exist.

Delegating comparable change authority to an autonomous agent does not create the underlying consequence; it compresses the time available to detect, understand and reverse an unsafe action. The evidentiary threshold for autonomous authority must therefore be higher, especially for actions capable of affecting emergency communications or other essential services.

The Optus precedent also presents an objection that this paper must meet: governance processes can exist and still be poorly applied. Continuous Shadow Deployment is not immune to misconfiguration or neglect. Its advantage is narrower. It places identity, permission checks, action-class limits, evidence capture, rate limits and rollback gates in the execution path, so that an unsupported action is technically rejected rather than merely prohibited by a procedure document. The architecture reduces dependence on procedural compliance; it does not eliminate implementation failure.

> **OPERATIONAL PRECEDENT**
> That was a firewall upgrade - not an autonomous agent.
> The incident establishes the human stakes of production change and the need for a higher evidence threshold when authority is delegated to systems acting at machine speed.

4. Predictive Twins as Preventive Governance

Network digital twins are developing from passive representations used for planning toward predictive systems connected to operational control. Elena Fersman has described a model in which a smart twin captures system dynamics, evaluates candidate actions before they reach the live network and compares predicted outcomes with observed results to improve future predictions. This is a significant form of preventive governance: it creates a pre-execution gate and a learning loop.[5]

Nokia's 2026 RAN Digital Twin work likewise illustrates high-fidelity simulation for AI-native network design and optimisation. A predictive twin estimates consequences before execution; Continuous Shadow Deployment asks whether the agent's use of authority remains acceptable under live operating conditions.[6]

Yet a predictive twin remains a model, constrained by data quality, coverage, update frequency and represented dependencies. It can estimate a candidate action's effect but cannot establish how an agent behaves across every live signal, permission request, tool failure or inter-agent exchange. Preventive governance and evidentiary governance answer different questions.

Governance layer	Temporal position	Core question	Principal evidence
Design and sandbox assurance	Before live exposure	Does the system satisfy defined requirements and threat scenarios?	Test plans, test logs, red-team findings and residual-risk decisions
Predictive-twin gate	Immediately before execution	What is the modelled consequence of this proposed action?	Twin state, predicted outcome, uncertainty and gate result
Shadow authorisation evidence	Before delegated authority	What would the agent decide under live conditions, and how often is that decision acceptable?	Decision-event records, divergence analysis and adjudication
Runtime governance	During authorised operation	Is the agent acting within identity, policy, scope and safety constraints?	Policy decisions, tool calls, actions, overrides and rollback events
Post-market learning	After deployment and incidents	What changed, failed or drifted, and what must be corrected?	Outcome monitoring, incident records, trend analysis and corrective action

5. The Missing Layer: Evidentiary Governance

Evidentiary governance is the capacity to demonstrate, through reproducible records, that an autonomous system's authority was justified and remains controlled. It does not depend on hidden chain-of-thought, which may be unavailable, misleading or inappropriate as a compliance artifact. The evidence should describe observable decisions and controls.

For each consequential event, the operator should be able to show the operational state, proposed or executed action, applicable policies and permissions, human intervention, and outcome or recovery. These facts enable review without assuming that a textual rationale faithfully represents internal computation.

Preventive and evidentiary governance

A predictive twin helps prevent an unsafe action by estimating consequences. Continuous Shadow Deployment helps determine whether the agent has earned authority by collecting evidence of its decisions under live conditions. Runtime controls keep authorised actions within scope. Post-market monitoring tests whether that authority remains justified.

6. Continuous Shadow Deployment

Continuous Shadow Deployment is a staged governance process in which an AI agent is progressively exposed to live operational conditions before, during and after it receives execution authority. Unlike conventional shadow testing, it shadows an authority claim rather than only an output. Because a stateful agent's second action may depend on the network response to its first, no single non-executing shadow phase can establish full behaviour. The lifecycle therefore combines observational shadowing, stateful twin or replay feedback, and bounded live canary execution. Shadow mode remains available as a fallback for new action classes, material changes, drift, incidents and periodic revalidation.

6.1 Phase 0 - Design and authority specification

Before live signals are routed to the agent, the operator defines the intended purpose, prohibited actions, data boundaries, tool set, identity, permission scope, autonomy level, escalation conditions, safety constraints and maximum tolerable blast radius. Each proposed action class is assigned a risk tier and a default human-approval requirement.

6.2 Phase 1 - Observational shadow ingestion

The agent receives production telemetry through a controlled, read-only or non-executing path. It may analyse alarms, query authorised read tools and generate proposed first actions, but the execution interface rejects or quarantines every change. This mode is strong evidence for diagnosis, tool selection, permission requests, policy compliance and the proposed first action. It cannot by itself establish how a multi-step agent would behave after observing the consequences of that action. Sensitive production data remain subject to security, privacy, minimisation, access-control and retention requirements; shadow mode is not a legal exemption.

6.3 Phase 2 - Stateful shadow and the feedback problem

A recommender can be shadowed cleanly because it receives an input and emits an output without changing the world it observes. An autonomous agent is different. Its action may alter network state, and that altered state becomes the input to its next decision. If live execution is blocked, the subsequent trajectory becomes counterfactual. Observational shadowing therefore systematically underrepresents multi-step, recovery and reinforcing-loop behaviour unless the agent receives a credible representation of execution feedback.

Evidence mode	Feedback source	What it can establish	Indicative authority ceiling
Observational shadow	Live inputs; no execution feedback	Diagnosis, tool and permission requests, policy compliance, proposed first action	A0-A1
Twin- or replay-mediated shadow	Simulated, replayed or isolated execution response returned to the agent	Multi-step plans, recovery choices and feedback-loop behaviour within represented conditions	A1-A2; not proof of live production behaviour
Bounded live canary	Real execution in a reversible, low-blast-radius scope	Actual closed-loop behaviour, rollback and human intervention under production feedback	Gate toward A3 within the tested action class

Every decision-event record should identify whether feedback was live-observed, simulated, replayed or synthesised. No authority claim should exceed the realism of the evidence used to justify it. A3 or higher authority should not be granted solely from non-executing shadow evidence for actions whose safety depends on multi-step or closed-loop behaviour.

6.4 Phase 3 - Comparison, triage and adjudication

Proposals are compared with human decisions, existing automation, policy requirements and the evidence available for the executed or simulated path. Human action is not automatic ground truth, and independent adjudication does not manufacture objective truth. When the human executes Y while the agent proposed X, the live outcome of X is unobserved. The counterfactual may be estimated through a twin, replay, historical analogue or expert review, but some divergences must remain classified as unresolved rather than being falsely labelled correct or incorrect. Recommended dispositions are: agent demonstrated preferable; human demonstrated preferable; both acceptable; both unacceptable; agent action plausible but counterfactually unverified; insufficient evidence; or controlled live evaluation required.

Evidence source	Primary use	Core limitation
Hard safety or policy invariant	Determine whether an action was prohibited, out of scope or inconsistent with an explicit rule	Rules may be incomplete, ambiguous or wrongly specified
Observed live outcome	Evaluate the action that was actually executed	Does not reveal the outcome of the rejected counterfactual

Evidence source	Primary use	Core limitation
Twin, replay or isolated execution	Estimate the unexecuted path and multi-step consequences	Dependent on fidelity, coverage and model error
Historical analogue	Compare with similar network states and prior actions	No two operational states are identical
Independent expert adjudication	Interpret mixed evidence and operational context	Relocates judgment; it does not create perfect ground truth
Unresolved counterfactual	Preserve uncertainty honestly and identify a need for further testing	May delay authority expansion

Every consequential event should be recorded, but not every event requires manual adjudication. Safety-critical, novel, low-confidence, policy-violating, high-blast-radius and emergency-service-affecting divergences should receive full review. Repetitive low-severity events can be handled through automated triage, stratified sampling and escalation thresholds. Periodic deep review of apparently successful events is still necessary to detect silent systematic error.

6.5 Phase 4 - Graduated authority

Authority is granted by network domain, action class, reversibility and risk tier. A system may be authorised to execute low-impact parameter changes while remaining in recommendation mode for routing, security isolation or emergency-service-affecting actions. The grant is time-bounded and linked to explicit metrics, change controls and named accountable owners.

6.6 Phase 5 - Runtime sentinel and reversible execution

During authorised operation, an independent policy and observability layer verifies agent identity, permission, action scope, current operating condition and cumulative change rate. Reversible actions create checkpoints. High-impact actions require stronger verification or human approval. Safety triggers can halt execution, roll back the action or return the system to shadow mode.

6.7 Phase 6 - Continuous revalidation

New model versions, prompts, tools, permissions, topology changes and policy changes re-enter shadow evaluation according to materiality. Drift, unexplained divergence, incident occurrence or deteriorating intervention performance can automatically reduce authority. The system therefore treats autonomy as a revocable licence rather than a permanent feature flag.

Phase	Evidence / execution mode	Primary output	Exit condition	Authority implication
0. Specification	No live exposure	Authority map, risk tiers and prohibited-action register	Design review and control readiness	No authority granted
1. Observational shadow	Live inputs; no execution feedback	First-action decision-event records	Coverage and record completeness	Supports A0-A1
2. Stateful shadow	Twin, replay or isolated feedback	Multi-step trajectories and recovery evidence	Fidelity and scenario limits documented	Supports A1-A2
3. Adjudication and triage	No live execution	Severity-weighted dispositions and unresolved counterfactuals	No unresolved critical cases for proposed scope	Gate to A2
4. Bounded live canary	Selected reversible production actions	Closed-loop execution, rollback and intervention evidence	Stable evidence within tested scope	Gate toward A3
5. Runtime sentinel	Authorised within explicit limits	Policy, action, override and rollback records	Continuous; authority expands or contracts	A3-A5 by separately evidenced scope
6. Revalidation	Shadow or reduced authority after material change	Change-specific evidence and renewed approval	Material-change gate satisfied	Contracts to any lower level, including A0

7. The Decision-Event Record and Draft Vendor-Neutral Schema

The decision-event record is the core evidence object and the most immediate candidate for telecom standardisation. It should be machine-readable, time-synchronised, tamper-evident and linked across the agent, model, tool, network and human-oversight layers. The table below is a draft vendor-neutral schema: it specifies observable facts needed to reconstruct conduct without exposing unnecessary subscriber data, protected credentials or unsupported claims about hidden reasoning.

Schema group and illustrative fields	Minimum content
Identity (event_id, agent_id, model_version, policy_version)	Agent identifier; model and orchestration version; policy version; deployment instance; authorised owner
Operational context (timestamp, domain, topology_ref, work_order_id)	Timestamp; network domain; relevant topology or state reference; incident or work-order identifier; data-quality flags
Input provenance (input_refs, telemetry_sources, freshness_flags)	Input-event references; telemetry sources; retrieval sources; transformations; missing or stale data indicators
Tools and permissions (tool_id, service_identity, requested_scope, policy_result)	Tools invoked; credential or service identity; requested and granted permission; policy decision; denied requests
Decision (proposed_action, action_class, risk_tier, constraints, uncertainty)	Proposed action; action class; risk tier; relevant constraints; confidence or uncertainty indicator; alternative or escalation
Human oversight (reviewer_id, displayed_context, intervention, response_time)	Reviewer or overseer; information displayed; approval, rejection, modification, override or no-intervention; response time
Execution (command_ref, target, checkpoint, rate_limit, rollback_status)	Command or API action; target; start and completion status; checkpoint; rollback capability; rate-limit status
Outcome (observed_effect, service_impact, incident_ref, disposition)	Observed network effect; service impact; policy deviation; incident linkage; corrective action and final disposition

Retention should be risk-based and aligned with applicable law, incident investigation, security and operational needs. Access to records should be restricted and monitored. Where a record contains personal or security-sensitive data, the operator should provide a minimised review view while preserving protected source evidence under controlled access.

8. Graduated Authority and Evidence Thresholds

Authority should not be granted through a single accuracy percentage. Telecom actions vary significantly in reversibility, time sensitivity and public consequence. A severity-weighted framework should combine performance, uncertainty, policy compliance, oversight effectiveness and recovery evidence.

Authority level	Permitted behaviour	Illustrative gate
A0 - Observe	Read live signals and create decision records; no recommendations served	Identity, data access and logging validated
A1 - Recommend	Recommendations shown to humans; no execution	Coverage achieved; critical-risk false negatives below threshold; reviewers can interpret output

Authority level	Permitted behaviour	Illustrative gate
A2 - Execute with approval	Agent prepares action; qualified human approves each action	Approval interface, policy checks, checkpoint and rollback demonstrated
A3 - Bounded autonomous execution	Agent executes specified reversible, low-to-moderate impact actions	Sustained evidence period; severity-weighted divergence within threshold; safe-stop and rollback service levels met
A4 - Domain autonomy	Agent coordinates authorised actions within one domain and objective set	Cross-condition validation; independent assurance; degraded-mode and multi-agent controls demonstrated
A5 - Cross-domain autonomy	Agent coordinates across domains under strategic objectives	Exceptional governance; formal risk acceptance; independent continuous monitoring and emergency control

The lifecycle phases and authority levels are related but not identical. Observational shadow generally supports A0-A1, stateful twin or replay evidence supports A1-A2, and bounded live canaries provide the production feedback needed to justify A3 within a specified action class. A4 and A5 require separate cross-condition, multi-agent and degraded-mode evidence; any material change can contract authority.

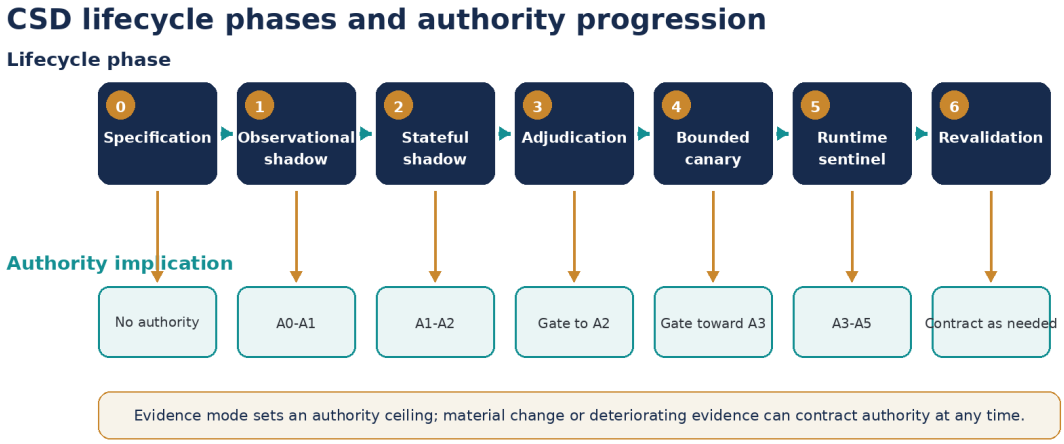


Figure 1. Continuous Shadow Deployment lifecycle phases and authority progression.

Illustrative evidence dimensions include: minimum observation duration; number and diversity of decisions; severity-weighted error rate; false-negative rate for safety-critical conditions; unauthorised-permission request rate; percentage of actions with complete records; escalation quality; human response time; rollback success; cumulative-change limits; performance during degraded telemetry; and absence of unresolved critical divergences. Operators should publish internal definitions and rationale rather than importing universal numbers without validation.

8.1 Divergence is a signal, not automatically an error

When an agent and human differ, the result may reveal agent failure, human error, ambiguous policy or a genuinely novel condition. A mature programme uses divergence to improve both system and operating procedure. The adjudication process should therefore record not only who was judged correct, but whether the existing rule, data or escalation pathway was adequate.

8.2 Authority decay and automatic contraction

Authority should contract when the environment materially changes or evidence deteriorates. Triggers may include a new software version, changed tool permissions, topology migration, repeated uncertainty, policy violations, delayed human response, rollback failure, a significant incident or declining record completeness. Contraction can be limited to one action class rather than disabling the entire system.

9. Human Oversight That Works at Network Speed

Human oversight is effective only when the overseer has competence, authority, information and time. Assigning a person to a dashboard does not create meaningful control if the person cannot understand the alert, determine the affected service, stop the action or recover the network before harm occurs.

Oversight measure	Operational question
Comprehension latency	How long does a qualified person need to understand the proposed or executed action and its likely consequence?
Intervention window	How much time exists between detection and the point at which intervention can still prevent or reduce harm?
Authority clarity	Can the overseer reject, modify, pause, roll back or reduce the agent's authority without organisational escalation?
Information sufficiency	Does the interface show system state, uncertainty, affected services, policy constraints and recovery options?
Safe-stop behaviour	Does interruption move the system and network to a defined safe state rather than simply terminating the agent process?
Automation-bias control	Are reviewers periodically tested, rotated and required to independently verify high-consequence recommendations?
Coverage and fatigue	Is oversight staffed for the volume, timing and complexity of autonomous actions, including overnight and incident conditions?

These measures are particularly relevant to Article 14 of the EU AI Act where a telecom system is classified as high-risk. Article 14 requires appropriate and proportionate oversight that enables natural persons to understand relevant capabilities and limitations, detect anomalies, remain aware of automation bias, interpret outputs, disregard or reverse outputs and interrupt operation through a safe-stop procedure. It does not prescribe Continuous Shadow Deployment, but an operator can use shadow and graduated-authority evidence to demonstrate whether these capabilities work in practice.[7]

10. Regulatory and Standards Relevance

10.1 EU AI Act

The EU AI Act establishes lifecycle requirements for high-risk AI systems, including risk management, data governance, technical documentation, logging, transparency to deployers, human oversight, accuracy, robustness, cybersecurity and post-market monitoring. Article 9 describes risk management as a continuous iterative lifecycle process and provides that testing may include testing in real-world conditions under Article 60. Article 60 creates a specific legal route for pre-market real-world testing with regulatory safeguards; it should not be confused with an operator's ordinary production shadow architecture.[7]

Classification remains use-case specific. Annex III point 2 applies to AI intended as a safety component in the management and operation of critical digital infrastructure. Telecom operations AI is not automatically high-risk because it is used by a telecom operator. The Commission's 2026 draft classification guidelines further illustrate the importance of intended purpose and the safety-component test. Providers and deployers also carry different obligations, and an operator may become a provider in defined circumstances such as substantial modification or change of intended purpose.[7, 9, 10]

The European Parliament approved the Digital Omnibus on AI on 16 June 2026, and the Council gave final approval on 29 June. The legislative act was signed on 8 July but, as of 18 July 2026, remained pending publication in the Official Journal; it will enter into force on the third day following publication. Under the adopted text, the application date for stand-alone Annex III high-risk systems is 2 December 2027, while the corresponding date for high-risk AI embedded in products under Annex I is 2 August 2028. This runway should be used to develop operational evidence rather than only policy documentation.[8, 20, 25]

Status note: Legislative status verified as at 18 July 2026. Readers should confirm subsequent Official Journal publication and entry into force before relying on this timeline.

10.2 Governance and standards crosswalk

Continuous Shadow Deployment can be mapped to existing instruments without being presented as a substitute. NIST AI RMF 1.0, its Critical Infrastructure Profile work, the AI Agent Standards Initiative and the Resource Center provide lifecycle risk outcomes and emerging implementation resources.[12, 14, 15, 19] CISA and international-partner guidance emphasises agent identity, least privilege, monitoring and secure integration with operational technology.[11, 17] GSMA and TM Forum work addresses telecom-specific agent use cases, coordination, benchmarking, zero-trust runtimes and Level 4 operations.[2, 3, 4, 13, 16, 18] The crosswalk below identifies the distinct operational contribution of each CSD component.

CSD component	NIST / critical infrastructure	EU AI Act relevance where in scope	Agentic security and telecom interoperability
Authority specification	GOVERN and MAP: purpose, context, roles and risk tolerance	Risk management, intended purpose and actor responsibilities	Identity, least privilege and comparable autonomy terminology
Decision-event record	MEASURE and MANAGE: evaluation, monitoring and incident evidence	Logging, technical documentation, oversight and post-market evidence	Accountability telemetry and a vendor-neutral event schema
Runtime policy gateway	MANAGE: treatment, response and control of risk	Human oversight, robustness, cybersecurity and safe interruption	Agent identity, permission enforcement, action approval and evidence export
Revalidation and authority contraction	Continuous lifecycle monitoring and corrective action	Post-market monitoring and response to material change	Drift monitoring, rollback and interoperable authority reduction

10.3 Cost, incentives and proportional assurance

CSD has real costs: duplicate inference, evidence storage, integration with tool gateways and change systems, twin or replay infrastructure, reviewer capacity and a slower path to full automation. The programme should therefore be risk-proportionate. Universal event logging does not require universal human review, and the most demanding evidence should attach to irreversible, security-sensitive, emergency-service-affecting or high-blast-radius actions.

The incentive is not only future compliance. A credible evidence layer reduces outage and liability exposure, improves incident reconstruction, strengthens vendor accountability, supports insurance and procurement assurance, and enables autonomy to expand without relying on undocumented confidence. Shadow phases can also reveal ambiguous procedures, excessive permissions, weak telemetry and recurring human decision errors before execution is enabled.

10.4 Emerging economies and vendor-embedded autonomy

In many emerging-economy markets, autonomous capability may arrive through vendor platforms, managed services, cloud orchestration or group-level operating models rather than in-house development. The operator may remain accountable while having limited control over model updates, orchestration logic, permissions, logging or evidence export.

Autonomy is therefore imported as a capability while assurance must be established locally. CSD should remain operator-controlled and vendor-independent: the operator should receive proposed actions through a non-executing interface, apply local authority thresholds, maintain its own decision-event record, contract authority automatically and export evidence even where the underlying model is vendor-controlled. Procurement contracts should guarantee the required technical access and audit rights.

ITU work already emphasises capacity building, repeated training, regional collaboration and standards participation for developing countries. For autonomous telecom operations, practical support could include shared evaluation facilities, common action-class test profiles, regulator training, pooled incident learning and a vendor-neutral decision-event schema.[23, 24]

A minimum viable implementation can begin with one high-consequence action class, universal event capture, risk-based sampling, local authority gates and vendor-provided evidence interfaces. Shared replay facilities and pooled specialist review can reduce cost. Operators should not have to choose between advanced autonomy without sufficient evidence and exclusion because full-scale assurance is unaffordable.

11. Three Governance Readiness Questions

These questions extend the three operational-sovereignty questions in ITS Issues Paper No. 1. The earlier questions address control over data, models, jurisdiction and decision reconstruction; the questions below address whether operational authority has been justified, bounded and made interruptible.

11.1 Can the operator reconstruct every consequential decision and action?

For any selected event, the operator should be able to identify the agent and version, relevant operational state, inputs and provenance, tools and permissions, proposed and executed action, applicable constraints, human intervention, observed result and recovery. A generic system description is not a substitute for an event-specific record.

11.2 Has authority been earned against pre-defined and independently adjudicated thresholds?

The operator should be able to show the evidence period, scenario and traffic coverage, severity-weighted divergence, unresolved cases, reviewer independence, performance under degraded conditions and the formal decision that granted each action class. Authority granted by project schedule, vendor assurance or executive preference is not evidentiary governance.

11.3 Can a qualified human intervene within the available intervention window and return the system to a safe state?

The operator should demonstrate—not merely assert—that oversight personnel receive intelligible information, possess authority, can pause or reverse actions, and can safely reduce autonomy. The intervention test must be conducted under realistic incident conditions, including high alert volume, incomplete telemetry and cross-domain effects.

12. Minimum Assurance Baseline and Recommendations

12.1 For operators

- Maintain an inventory and authority map for every autonomous or embedded vendor agent, covering versions, identities, permissions, action classes, accountable owners and prohibited actions.
- Use short-lived identities, least privilege, independent runtime policy enforcement and a tamper-evident evidence layer that the assessed agent cannot edit or suppress.
- Match the evidence mode to the authority claim: observational shadow for first actions, twin or replay feedback for multi-step evaluation, and bounded live canaries before closed-loop A3 authority.
- Record every consequential event; apply full review to high-severity, novel or policy-violating divergences and risk-based triage to the long tail. Grant and revoke authority by action class, domain, reversibility and consequence, with checkpoints, rollback and automatic contraction.
- Test human comprehension, intervention, safe-stop and recovery under realistic outage and degraded-telemetry conditions, and integrate the evidence with change management, cybersecurity, incident reporting and post-market monitoring.

12.2 For national telecom and critical-infrastructure regulators

- Classify autonomous action classes by service, safety, security and rights impact, and establish protected incident-sharing mechanisms that preserve sensitive topology and subscriber data.

- For high-consequence action classes, require evidence from an appropriate live-condition mode - including bounded canary execution where safety depends on feedback loops - before autonomous execution is authorised.
- Specify minimum event-record fields and evidence-quality labels, and require operators to demonstrate safe interruption, rollback and authority contraction in realistic conditions while protecting personal data, network security and legitimate trade secrets.

12.3 For standards bodies and international organisations

- Develop a vendor-neutral, machine-readable decision-event schema and common autonomy, authority-gate and evidence-mode labels that map to NIST, EU, ITU, TM Forum, GSMA and operator controls.
- Define interoperable interfaces and test profiles for identity, permissions, action approval, feedback-loop behaviour, rollback, evidence export and multi-agent interference.

13. Research Agenda and Call for Collaboration

The next stage is empirical validation. ITS invites operators, NOC and service-assurance practitioners, vendors, regulators, standards bodies, researchers and civil-society organisations to test the framework, identify counterexamples and contribute evidence on the following questions:

1. Which action classes and minimum evidence thresholds are appropriate for observational shadow, stateful twin or replay evaluation, and bounded live canary execution - particularly where operators have limited internal assurance capacity?
2. Which combinations of predictive-twin fidelity, historical replay, expert adjudication, automated triage, sampling and escalation can evaluate multi-step feedback at scale without hiding rare high-consequence failures?
3. Which common event schema, privacy controls, vendor interfaces and shared regional assurance resources can make CSD practical across multi-operator, cloud, managed-service and emerging-economy environments?

Substantive responses will inform the next ITS technical specification and the design of proposed operator or regulator pilots. Contributors may also be invited to participate in structured workshops or validation exercises.

14. Conclusion

Telecom operators are beginning to delegate operational authority to systems that can perceive, plan, call tools and act. The Optus Triple Zero outage shows why this distinction matters: a routine firewall-upgrade change within a human-governed process disrupted emergency calling for more than 14 hours. Autonomous operation brings the same public stakes while compressing the time available to recognise and reverse error.

Predictive twins provide a preventive layer by estimating candidate actions. CSD adds evidence about an agent's claim to use tools, permissions and operational authority under real conditions. Because non-executing shadow cannot reproduce stateful feedback, credible assurance combines observational shadowing, twin or replay feedback and bounded live canaries, followed by independent runtime policy, tamper-evident observability, human intervention and automatic authority decay.

The core proposition is simple: autonomous authority should be earned through evidence, granted in bounded increments and remain continuously revocable. Operators that can reconstruct decisions, justify authority and demonstrate safe intervention have begun to govern agentic AI operationally. The network should never move faster than its accountability.

About the Author

Gaurav Arora is the founder of the Institute for Technology Stewardship and a telecom and critical-infrastructure practitioner with more than two decades of experience in network operations, service delivery, technical leadership and technology programme management. His research focuses on translating operational practice into governance mechanisms for autonomous systems.

About the Institute for Technology Stewardship

The Institute for Technology Stewardship (ITS) is an independent research initiative focused on the governance of autonomous and emerging technology systems. ITS bridges operational expertise with policy research, developing practical frameworks intended to keep technology deployment in critical infrastructure safe, transparent, accountable and human-centred.

ITS works from the premise that institutions with long experience of human-machine collaboration at scale - including telecom network operations, air traffic control and other critical infrastructure - hold lessons that should be more fully represented in AI governance and international policy discussions.

Declarations

Funding: No external funding was received. **Competing interests:** The author declares no competing interests. **Data availability:** No original dataset was generated; the analysis is based on publicly available sources. **Ethical approval:** Not applicable.

Suggested Citation

Arora, Gaurav. 2026. Beyond the Sandbox: Evidence-Based Governance for Autonomous AI Agents in Telecom Networks. Institute for Technology Stewardship, Issues Paper No. 2, July 2026.

Disclaimer

This issues paper provides independent policy and operational analysis and does not constitute legal advice. It does not represent the position of any operator, regulator, standards body, company or government. References to organisations and initiatives are included for analytical purposes and do not imply endorsement or criticism. Legal classification and obligations should be verified for the jurisdiction, role, intended purpose and date of any actual deployment.

References

1. Deutsche Telekom. Dt. Telekom and Google Cloud: AI-Agent MINDR for Superior Network Quality, 25 February 2026. Reports that the production RAN Guardian Agent autonomously triggered more than 100 remediation actions and describes the broader MINDR architecture; the public source does not detail the approval workflow for each action. Source
2. GSMA. Agentic AI in Telecom, 3 September 2025. Industry overview of agentic AI use cases and autonomous network operation. Source
3. TM Forum. Zero-Trust Agents for Autonomous Networks, Catalyst C26.0.933, 2026. Describes identity, policy, observability and evidence for governed agent runtimes. Source
4. TM Forum. Autonomous Networks Level Solution Packs, 2026. Includes Level 4 fault-management and quality-optimisation solution packs. Source
5. Ericsson. Five Ways a Network Digital Twin Enables Safe Autonomy, 16 April 2026. Elena Fersman on predictive and interactive twins, pre-execution evaluation and learning loops. Source
6. Nokia. Nokia Launches Nokia RAN Digital Twin to Turbo-Charge AI-Native 6G, Powered by NVIDIA Aerial Omniverse Digital Twin, 26 February 2026. Source
7. European Union. Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence, 13 June 2024. See Articles 6, 9, 12-15, 25-26, 60-61 and 72, and Annex III. Source
8. Council of the European Union. Artificial Intelligence: Council Gives Final Green Light to Simplify and Streamline Rules, 29 June 2026. Confirms new application dates for high-risk AI requirements. Source
9. European Commission. Draft Commission Guidelines on the Classification of High-Risk AI Systems, 19 May 2026. Draft guidance and practical examples concerning Article 6 and Annex III. Source

10. European Commission AI Act Service Desk. Article 25: Responsibilities Along the AI Value Chain, 2026. Official explanatory access to the AI Act text and actor-role provisions. Source

11. CISA, NSA and International Partners. Careful Adoption of Agentic Artificial Intelligence Services, 1 May 2026. Joint non-binding guidance on agent identity, privilege, interactions and monitoring. Source

12. NIST. AI Agent Standards Initiative, 17 February 2026. Programme addressing standards, security and interoperability for AI agents. Source

13. TM Forum. Agent Fabric: A2A-T Runtime - Phase III, Catalyst C26.0.910, 2026. Multi-vendor runtime governance, visibility and trust scoring. Source

14. NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, January 2023. Source

15. NIST. Concept Note: AI RMF Profile on Trustworthy AI in Critical Infrastructure, 7 April 2026. Profile development for AI-enabled capabilities across critical infrastructure. Source

16. GSMA. Coordinating Agentic AI for Autonomous Networks, 18 December 2025. Example of central coordination across specialist agents. Source

17. CISA and International Partners. Principles for the Secure Integration of Artificial Intelligence in Operational Technology, 3 December 2025. Non-binding guidance for safe and secure AI integration in operational technology. Source

18. GSMA. Open-Telco LLM Benchmarks Launches to Advance AI in Telecoms, 25 February 2025. Demonstrates domain-specific model evaluation challenges in telecommunications. Source

19. NIST. AI Risk Management Framework Resource Center, Updated 2026. Includes AI RMF updates and critical-infrastructure profile activity. Source

20. European Parliament. AI Act: EP Approves Simplification Measures and “Nudifier” App Ban, 16 June 2026. Records Parliament’s final approval, vote and revised application dates. Source

21. Independent Review of the Triple Zero Outage at Optus on 18 September 2025. Final report on the cause, duration, failed emergency calls, fatalities, change-management failures and recommendations. Source

22. Australian Communications and Media Authority. ACMA Statement on Optus Triple Zero Investigation, 22 September 2025. Official regulator statement on emergency-call obligations and investigation scope. Source

23. International Telecommunication Union. AI Standards for Global Impact: From Governance to Action, 2025. Calls for capacity building and support for developing-country policymakers and regulators, including in Africa and Latin America. Source

24. International Telecommunication Union Regional Office for Asia and the Pacific. Workshop on AI Standards for Increasing the Efficiency of Telecommunications/ICTs: Shaping the Future Responsibly - South Asia Report, 2025. Documents demand for repeated training, regional collaboration and South-South cooperation on AI and telecom standards. Source

25. Council of the European Union. PE-CONS 30/26, Regulation Amending the AI Act (Digital Omnibus on AI), 18 June 2026. Adopted legislative text, including entry into force following Official Journal publication. Source