

Reality Drift Governance Readiness Audit

Evaluating Drift Detection, Semantic Fidelity, and Corrective Capacity

A. Jacobs

Purpose

This audit determines whether an organization can detect and correct AI drift after deployment.

Policies, approvals, documentation, and monitoring are necessary, but they do not establish readiness by themselves. An organization is ready only when evidence from real-world outcomes can invalidate its representations, interrupt operation, and change what happens next.

The audit examines four capacities: orientation, representational integrity, detection, and correction. It requires evidence of exercised authority rather than the existence of procedures alone.

Core Definitions

Reality Drift: The gradual loss of alignment between a system's internal representations and the reality those representations were created to describe, govern, or support. A drifting system may remain operational, coherent, and apparently successful even as its capacity for reality-based correction weakens.

Semantic Fidelity: The degree to which a representation preserves the meaning, intent, context, distinctions, and referential relationships necessary for its intended use.

Constraint Collapse: Occurs when feedback, consequence, and correction remain present but lose their binding force. Signals continue circulating, yet they no longer compel change.

Readiness Standard

The central governance question is:

Can the organization detect and correct drift before it becomes embedded in decisions, policies, workflows, or institutional behavior?

A monitoring system is not corrective merely because it records problems. Feedback becomes corrective only when disconfirming information can materially revise the representation, workflow, decision rule, deployment scope, or continued use of the system.

A. Orientation

1. Purpose and Ground Condition

The organization can state the system's intended users, supported task, decision context, workflow boundaries, required human judgment, prohibited uses, and real-world outcome. Broad claims such as improving productivity or enhancing decision making are insufficient.

2. Success and Failure Conditions

The organization defines success using outcomes that remain connected to the purpose. It also defines unacceptable failure, escalation thresholds, and conditions under which the system must lose authority.

B. Representational Integrity

3. Proxy Validity

Operational measures, benchmarks, ratings, and completion indicators are treated as proxies. Their relationship to the underlying purpose is periodically tested and can be revoked.

4. Semantic Fidelity

Reviews test whether intent, context, distinctions, uncertainty, tone, and practical usefulness survive retrieval, summarization, classification, generation, and handoff.

5. Provenance and Traceability

Important outputs can be traced through their source materials and transformations. The organization distinguishes direct observation, interpretation, summary, generated material, and recursive reuse.

6. Workflow Mapping

The organization knows where AI outputs enter the process, who uses them, which decisions they influence, how users adapt around them, and whether outputs return as future inputs.

C. Detection

7. Longitudinal Evaluation

The organization preserves stable evaluation cases, current real-world cases, version records, qualitative review, and comparison over time. It does not interpret stable use or low complaint volume as proof of alignment.

8. Observability

Inputs, outputs, interactions, transformations, recurring failures, downstream use, and affected groups can be inspected. Usage tracking alone does not qualify as observability.

9. Independent Grounding

At least some evaluation uses evidence that is independent of the representations and incentives being tested. Dashboards do not validate dashboards. Synthetic judges do not serve as the only judges of synthetic outputs.

D. Correction

10. Corrective Authority

Named people can change prompts, retrieval, workflow, access, deployment scope, evaluation criteria, or continued use. Reviewers who can only recommend changes do not constitute a complete corrective mechanism.

11. Escalation and Stopping Conditions

Material failures can trigger restriction, rollback, pause, or retirement. The conditions and responsible owner are documented before an incident occurs.

12. Reality Recapture and Revalidation

The organization periodically returns to underlying conditions, gathers evidence not already filtered through the current system, and revalidates proxies, categories, and assumptions.

13. Vendor and Change Control

Vendor systems are tested against actual organizational use. Model updates, policy changes, routing changes, new data sources, new users, and new downstream uses trigger reassessment.

Evidence Standard

Do not award readiness based only on policy statements. For each domain, request an artifact or demonstrated example:

- A defined owner with decision authority
- A current workflow or dependency map
- A completed drift review using preserved cases
- A trace from output to source and transformation history
- A case where feedback changed deployment, scope, evaluation, or continued use
- A tested rollback, pause, or retirement procedure
- A proxy revalidation against an independent outcome

Readiness Scorecard

Assess each domain across three levels of readiness.

Purpose

Not ready: Purpose is vague or aspirational.

Partially ready: Purpose is documented but weakly tied to outcomes.

Ready: Users, task, boundaries, outcome, and failure conditions are explicit.

Proxy Validity

Not ready: Internal measures are treated as success.

Partially ready: Some proxy review occurs.

Ready: Proxies are independently tested and can lose authority.

Semantic Fidelity

Not ready: Outputs are judged mainly by fluency or correctness.

Partially ready: Qualitative review exists.

Ready: Meaning, intent, context, distinctions, and usefulness are evaluated.

Provenance

Not ready: Sources and transformations cannot be reconstructed.

Partially ready: Some sources are retained.

Ready: Material outputs have reconstructable transformation histories.

Workflow

Not ready: AI is added without dependency mapping.

Partially ready: Affected processes are partly mapped.

Ready: Decision points, downstream effects, and recursive reuse are documented.

Monitoring

Not ready: Usage or incidents are tracked.

Partially ready: Output samples are reviewed irregularly.

Ready: Stable and current cases are evaluated longitudinally.

Grounding

Not ready: Internal checks validate internal outputs.

Partially ready: Some external evidence is used.

Ready: Independent outcomes can contradict and revise the system.

Authority

Not ready: Reviewers can report but not compel change.

Partially ready: Escalation exists but authority is unclear.

Ready: Named owners can restrict, roll back, pause, or retire.

Post-Deployment

Not ready: Approval ends at launch.

Partially ready: Reviews occur irregularly.

Ready: Scheduled and trigger-based reassessment continues through retirement.

Critical Governance Tests

1. Can the system satisfy internal success criteria while the underlying condition deteriorates?
2. Who can interrupt, restrict, roll back, or retire the system when external evidence contradicts internal performance?
3. What evidence would cause the organization to conclude that its current representation is wrong?
4. Can frontline or affected users introduce evidence that changes categories, incentives, or decisions?
5. Can the source and transformation history of consequential outputs be reconstructed?
6. Has feedback ever forced a material change, or does it merely circulate?

Audit Output

Conclude the audit with six items:

- Overall readiness level
- Highest-risk representational layer
- Most authoritative unvalidated proxy
- Weakest corrective constraint
- Required stopping condition and authorized owner
- Next reality recapture and proxy revalidation date

Decision Rule

An organization is not ready when it can observe problems but cannot make them binding.

Extensive documentation, monitoring, and review can coexist with Constraint Collapse. Governance becomes real when evidence from reality can invalidate the representation and alter the system's continued operation.

The absence of failure is not proof of alignment. Coherence is not contact.