

Preserving Meaning Across Systems

Semantic Fidelity and the Limits of Representation

A. Jacobs

Purpose

The Reality Drift framework examines how systems lose alignment with the conditions they were created to describe, govern, or support while continuing to function. Its central concern is the growing authority of representations after their relationship to the underlying world has weakened.

Systems coordinate through records, scores, indicators, and summaries that make complex conditions easier to act upon. As these representations move into decisions and subsequent workflows, what they omit can become unavailable to the people and systems relying on them. A summary may eventually serve as a planning premise even though the qualifications in its source would have changed the decision.

This design brief explains how representations change as they move through systems, how to recognize consequential losses of meaning, and what allows evidence to correct the decisions built from them. It brings the framework's diagnostic vocabulary, Representation Stack model, and transformation-to-failure pathways together as a basis for considering a semantic integrity layer across systems.

Coherence Without Contact

Reality Drift develops when a system's dependence on representations, divergence from underlying conditions, weakened correction, and continued operation reinforce one another over time. A drifting system can keep producing the evidence by which it is judged successful while the underlying need becomes harder to satisfy. Completed reports, followed procedures, and passing benchmarks may conceal the deterioration because they establish success within the system's own terms.

Ordinary, useful operations can contribute to this condition. Summarization can remove qualifications, metrics can flatten consequential differences, and optimization can reward a proxy that has become less reliable. When the omitted conditions can no longer constrain later decisions, local improvements can sustain broader misalignment.

The Drift Principle describes the underlying imbalance. Representations can be compressed, optimized, distributed, and reused faster than reality can correct them. As their circulation accelerates, later decisions increasingly depend on accounts whose originating conditions are difficult to recover. The framework follows how these representations change, acquire authority, and remain open to correction. The decisive question is whether evidence from the underlying condition can still change what the system does. Coherence is not contact, and the absence of failure is not proof of alignment.

A Vocabulary for Inspecting Transformation

The framework’s diagnostic vocabulary makes it possible to describe how systems lose alignment, why their representations remain convincing, and who bears the burden of correcting them. The following concepts establish the wider problem developed in this brief:

Concept	What It Identifies
Reality Drift	<i>A system remains operational and coherent while progressively losing corrective contact with the conditions it was created to describe, govern, or support.</i>
Drift Principle	<i>Representations can be compressed, optimized, distributed, and reused faster than reality can correct them.</i>
Semantic Fidelity	<i>The degree to which meaning, intent, context, and consequential relationships survive transformation.</i>
Constraint Collapse	<i>Feedback, consequence, and correction remain present but lose the binding force needed to compel revision or interrupt misaligned operation.</i>
Survivable Wrongness	<i>Misalignment that persists because it remains compatible with the continued operation of the system producing it.</i>
Synthetic Realness	<i>An optimized representation conveys stronger signals of credibility, completeness, or authenticity than the less legible reality it represents.</i>
Filter Fatigue	<i>The cognitive burden placed on individuals who must repeatedly establish what is relevant, original, and sufficiently grounded in representation-heavy environments.</i>

These concepts describe how systems lose corrective contact, what happens to meaning, why misalignment persists, and how its costs reach individuals. They provide a language for examining failures that can remain hidden behind coherent outputs and successful operation.

How Reality Becomes Representation

The Representation Stack describes how an underlying condition becomes something a system can measure, act upon, and explain:

Reality → Measurement → Metrics → Operational Representation → Optimization → Narrative

Measurement selects observable features from a condition, which metrics organize into comparable indicators. Those indicators enter models, classifications, dashboards, and targets that guide decisions and define what gets optimized. The resulting outcomes are interpreted through narratives that explain what happened and why.

As information moves through these layers, the basis for interpretation changes. Comparability may come at the expense of differences that matter locally, while a classification can make an ambiguous situation appear settled. By the time an outcome reaches a report or explanation, the selections that shaped it may be difficult to reconstruct. The Stack helps locate those transitions, including in workflows that skip, combine, or repeatedly revisit layers.

Once representations guide decisions, they also influence the conditions being represented. Organizations adapt their behavior to rankings and performance indicators, changing what subsequent measurements capture. The resulting observations can then appear to validate the representations that helped produce them:

Reality → Capture → Representation → Operational Proxy → Optimization → Behavioral Adaptation → Altered Reality → Recapture → Apparent Confirmation

Evaluating this loop requires examining how the system influenced the evidence it now treats as confirmation. Agreement between a model and an observed outcome may reflect adaptation to the model's categories or incentives. Representational inversion describes this development: the conditions being observed are increasingly reorganized to satisfy the representations originally created to describe them.

From Transformation to Failure

The transformation-and-failure-mode series examines how ordinary operations can develop into failures that remain compatible with continued operation. Survivable Wrongness describes this broader condition of persistence. Each pairing traces what gets lost during a transformation and how the resulting representation remains credible or usable as corrective contact weakens. These are pathways to examine rather than inevitable outcomes. Where a named process also supports useful operation, the description identifies how it develops when corrective contact deteriorates.

Transformation Category	Failure Mode	How the Failure Develops
Scaling and Optimization	Constraint Collapse	<i>Expansion outpaces the constraints capable of forcing revision or stopping misaligned activity.</i>
Citation and Recirculation	Evidence Laundering	<i>Repeated references to a shared origin acquire the appearance of independent corroboration.</i>
Proxy Optimization	Optimization Trap	<i>Improvement against the proxy reinforces behavior that moves the system farther from its underlying purpose.</i>
Proceduralization	Procedural Substitution	<i>Completing an authorized process becomes sufficient evidence that the underlying problem was resolved.</i>
Copying and Reuse	Provenance Failure	<i>Content travels without enough source and transformation history to reconstruct or challenge it.</i>
Recursive Compression	Recursive Mediation	<i>Successive representations become the basis for later understanding while access to originating conditions weakens.</i>
Behavioral Adaptation	Representational Inversion	<i>People reshape conditions around a representation, and the resulting conditions appear to confirm it.</i>
Synthesis and Pattern Fitting	Synthetic Coherence	<i>Fragments are assembled into a convincing account whose internal consistency exceeds its evidentiary support.</i>

Several pathways can reinforce one another within the same system. A claim may lose its origin through copying and reuse, gain apparent authority through citation, and enter a synthesis that conceals the missing evidence. Examining these connections helps explain how an initially limited distortion becomes an accepted basis for decisions and survives the processes intended to correct it.

When Representations Govern

Representational failures acquire wider consequences when the affected accounts determine funding, access, evaluation, or institutional action. Goodhart's Law, reward hacking, specification gaming, and goal displacement examine different ways that measurable or formal success can separate from intended purpose. Reality Drift follows the broader condition in which metrics, procedures, categories, and narratives continue governing decisions while the conditions they represent lose corrective authority.

Within institutions, that separation can appear through several connected pathways. A mission can remain prominent while losing its power to constrain decisions, and an authorized process can reach completion while the underlying problem remains unresolved. Explanations may absorb contradictory events without changing their central assumptions, while reporting cycles register success before longer-term consequences become visible. These pathways describe how purpose, procedure, interpretation, and timing can become misaligned within the same operating system.

A semantic integrity layer would therefore need to connect preservation of meaning with the arrangements through which representations acquire authority. Detecting a distortion has limited effect when incentives reward its continued use or reviewers lack the authority to revise it. The question extends from whether a representation remains faithful to whether the organization permits contrary evidence to change what governs its decisions.

Preserving Meaning Across the Chain

Semantic Fidelity concerns whether meaning, intent, context, distinctions, and referential relationships survive representation and transformation. What needs to survive depends on the representation's intended use. A summary prepared to locate a report may require less detail than an account used to justify a consequential decision.

The relevant test concerns what an omission changes. Restoring uncertainty may alter whether action is justified, while restoring an exception may determine whether a rule applies at all. These relationships can be lost even when an output retains recognizable facts and language from its source. They can also be distorted through emphasis or placement, as when an account includes every qualification but presents a conclusion those qualifications undermine.

Generative reconstruction can conceal the losses introduced by compression. Missing context is replaced with plausible connections and fluent explanation, giving the result a completeness the retained evidence does not support. Later systems inherit an apparently finished account with few visible indications of what disappeared. Fidelity evaluation therefore needs to examine the transformation history as well as the quality of the reconstructed output.

Following fidelity across a workflow expands the scope of evaluation. A source’s uncertainty must survive retrieval, interpretation, summarization, storage, and exchange if it is to constrain a later decision. Each stage can perform its assigned function successfully while changing the meaning inherited by the next. Evaluating only the final output, or checking each component against its own requirements, can leave those changes unexamined.

A faithful account can also reproduce a mistaken or outdated source. The chain must preserve enough meaning to remain answerable to its origins while allowing new evidence to challenge those origins and revise dependent decisions. Preserving meaning therefore requires coordination across functions usually handled separately. Retrieval, provenance, semantic modeling, memory, interoperability, and correction all participate in maintaining the same relationship between what a representation means and what it is being used to do.

Toward a Semantic Integrity Layer

Retrieval, provenance, ontologies, knowledge graphs, memory, interoperability, and evaluation each address part of this chain. Their shared concern becomes visible when a representation is followed across them. What retrieval selects is interpreted through categories and relationships, transformed into an answer, retained in memory, exchanged between systems, and eventually used to guide action. Meaning can change anywhere along that path, while responsibility for preserving it remains divided among the systems handling each part.

The proposed semantic integrity layer would connect these responsibilities around the preservation of meaning and corrective contact. Its organizing concern is whether a representation remains answerable to its source, purpose, and underlying conditions throughout its use. Retrieval determines what evidence can enter that relationship, provenance makes its history recoverable, and interoperability determines what survives when it crosses systems. Their contributions become parts of a continuing task of preserving meaning and corrective contact.

The following domains show where responsibility for semantic integrity would need to be maintained.

Domain	Responsibility Within the Layer
Retrieval and Relevance	<i>Make task-relevant evidence available, including qualifications and contradictory material.</i>
Provenance and Lineage	<i>Preserve origins, transformation history, and dependencies so representations can be traced and challenged.</i>
Ontologies and Knowledge Graphs	<i>Make concepts and relationships explicit, including connections among claims, sources, and decisions.</i>
Interpretation and Transformation	<i>Preserve consequential relationships through reasoning, summarization, translation, classification, and synthesis, while making changes inspectable.</i>
Memory and Agency	<i>Carry context and constraints across time so later plans and actions remain answerable to originating objectives.</i>
Interoperability	<i>Preserve meaning, context, constraints, and provenance across systems with different schemas and assumptions.</i>

Domain	Responsibility Within the Layer
Evaluation and Correction	<i>Make consequential changes detectable and connect findings to the authority needed to revise or interrupt downstream action.</i>

A representation could carry its originating evidence, conditions of validity, dependencies, uncertainty, and transformation history alongside its content. It could also identify circumstances that would require revision and a route for correction. Receiving systems would then have a basis for examining what they inherited and whether it remains suitable for a new use.

That history would need to expose changes in meaning, including added assumptions, altered relationships, and omitted constraints, while identifying which version informed a decision. Where compression prevents reconstruction, that limit should remain visible. Dependencies between sources would also need to distinguish independent observations from repeated interpretations of a shared origin, including generated accounts that later return through retrieval as apparent external support.

The semantic integrity layer could operate across existing systems and interfaces, connecting their responsibilities throughout a representation’s use. A change in meaning detected in one part of the workflow would need to remain visible wherever that representation is reused, with a route for correction to reach dependent decisions. The framework provides the language and models for examining that continuity and identifying where it breaks.

The layer would itself need corrective contact. Its judgments would require testing against independent observations and outcomes, including evidence from people affected by the system. Responsibility for acting on that evidence would need to be explicit, with authority to revise categories, change evaluation criteria, restrict use, or stop operation when continued use is no longer justified.

A Short Example

Consider a fictional report:

“In this small observational sample, Clinic A had fewer repeat visits after introducing remote follow-up. Patients also received additional staffing support. The report cannot determine which change produced the difference, and the result should not be generalized to other clinics without further study.”

An AI summary states that remote follow-up reduced repeat visits at Clinic A. A later answer retrieves the report, cites it, and recommends adoption at another clinic. That recommendation enters a planning brief which subsequently becomes context for another assistant. The relevant source was found, but its causal limits did not govern the interpretation. Compression removed the staffing change and uncertainty, allowing a local observation to become a causal finding. Reuse then carried that finding into a recommendation elsewhere, while the citation gave later readers a reason to regard it as grounded. Following the case across domains shows what the semantic integrity layer would connect:

Part of the workflow	What needs to remain visible
Retrieval	<i>Whether the passages containing the staffing change and causal limitation entered usable context.</i>
Interpretation	<i>Where an observed sequence became a causal conclusion.</i>
Transformation	<i>Which qualifications disappeared between report, summary, answer, and brief.</i>
Provenance	<i>How the recommendation and planning premise depend on the original report.</i>
Interoperability	<i>Whether limits attached to the finding survived transfer into another tool or system.</i>
Memory and agency	<i>Whether later reasoning treats the recommendation as an established fact or a conditional interpretation.</i>
Evaluation and correction	<i>Whether restoring the missing information changes the answer, brief, and resulting plan.</i>

A faithful concise account could say: “Repeat visits fell at one clinic after remote follow-up and additional staffing were introduced; the report could not establish the cause.” The report’s uncertainty then remains available to shape the decision rather than disappear during compression. The example also distinguishes source presence from source force. The report has presence when it is retrieved and cited. Its force becomes visible when its limitations change the answer and constrain the recommendation. If the answer is corrected while the planning brief continues circulating unchanged, the correction has reached only part of the chain.

A research program could follow a source through repeated summarization, translation, storage, retrieval, and agent handoff, comparing each stage with the originating material and intended use. It would examine where consequential relationships change, whether those changes affect decisions, and whether restoring lost information corrects dependent outputs. Such experiments would test where losses accumulate and which interventions preserve meaning and corrective contact.

Keeping Representations Answerable to Reality

Representations can acquire authority through their movement between systems, as a summary becomes evidence for a recommendation that enters a plan and eventually shapes action. The conditions that justified the original account can disappear along the way, even as each successive version becomes easier to use. Systems that accelerate this movement need a corresponding capacity to preserve meaning and carry correction through the consequences of earlier outputs.

The semantic integrity layer gives that capacity a common purpose across domains. Preserving a qualification through reuse and carrying a correction into dependent decisions become connected responsibilities, allowing evidence to challenge the assumptions governing subsequent action.

The Reality Drift framework makes this a problem that can be named, traced, and designed around. Its larger direction is toward systems whose growing capacity to transform representations is matched by a capacity to remain answerable to what they represent. As those representations gain the power to shape the world, the world must retain the power to correct them.