Hallucination, Grounding, Faithfulness, and Reality Drift

Why AI Systems Can Remain Coherent While Losing Contact With Reality

AI Systems and Reality Drift - Note #4 v1

A. Jacobs - Semantic Fidelity Project

The Basic Pattern

Modern AI systems often appear remarkably capable. They answer questions, summarize information, write code, retrieve knowledge, and support decisions that previously required human judgment.

At first, these systems can seem closely aligned with their intended purpose. Their outputs are useful, benchmark scores improve, users report successful interactions, and performance metrics rise. But strong performance can obscure a more fragile relationship between the system and the world it is being asked to represent.

The danger is not only that AI systems make mistakes. It is that they can remain convincing after their grounding has already weakened. A model may sound fluent, confident, and useful even when it is operating beyond the conditions its tests actually captured. In deployment, the failure may not appear as a crash or obvious malfunction. It may appear as coherent output that is no longer adequately constrained by evidence, context, or real-world feedback.

Different communities describe this challenge using different terminology. It appears in discussions of hallucination, grounding failure, faithfulness failure, distribution shift, benchmark overfitting, alignment failure, and model collapse. These are not identical problems, but they often reveal a shared structural risk: the system can keep functioning while the relationship between representation and reality becomes less reliable.

Coherence Is Not Contact

AI systems do not interact directly with reality. They interact with representations of reality. Training data represents the world, benchmarks represent performance, evaluation metrics represent capability, embeddings represent meaning, and retrieved documents represent knowledge.

These representations make AI systems usable at scale, but they also introduce distance. The system learns from compressed traces of the world rather than from the world itself. When those traces remain accurate enough, the system can appear reliable. When they become stale, incomplete,

distorted, or poorly matched to deployment conditions, the same system may continue producing fluent outputs without preserving the relationships that made those outputs trustworthy.

This is the key distinction. Coherence is an internal property of the output. Contact is a relationship between the output and the conditions it claims to describe. An AI system can preserve the first while gradually losing the second.

How AI Drift Emerges

Stage 1 — Representation

A model is trained using data that reasonably reflects the environment it is intended to operate within. At this stage, representations remain connected to reality, and performance appears reliable.

Stage 2 — Optimization

Training improves, benchmarks improve, and capabilities expand. The model becomes increasingly effective at generating outputs that satisfy evaluation criteria.

Stage 3 — Decoupling

The environment changes, data distributions shift, new contexts emerge, and users ask questions not well represented by prior conditions. Evaluation systems may fail to capture these emerging weaknesses, and the model increasingly relies on patterns that no longer reflect the current environment.

Stage 4 — Drift

Outputs remain coherent, benchmarks may remain strong, and the system continues operating successfully. Yet fidelity to reality gradually weakens as the model increasingly reflects its internal optimization environment rather than the conditions it was designed to understand.

The absence of failure is not proof of alignment. It may only mean the system has stopped detecting the gap.

Semantic Fidelity and Constraint Collapse

At this point, the system has not necessarily failed on its own terms. What weakens is the structural relationship between the model's representations and the reality those representations are supposed to preserve.

An output may retain the surface appearance of understanding while losing the conditions, relationships, and constraints that made the response meaningful. This is where hallucination, grounding failure, and faithfulness failure become semantic fidelity problems. The issue is not only

whether the output is fluent or plausible. The issue is whether the representation still preserves the relevant relationship to evidence, source material, context, and real-world conditions.

AI drift also involves constraint collapse. Grounding mechanisms are supposed to keep model behavior answerable to evidence, context, objectives, or real-world feedback. When those constraints weaken, the model may still operate smoothly, but its outputs lose some of the corrective pressure that should keep them tied to reality.

The Representation Stack

AI drift becomes more serious when outputs move through a wider representation stack. Reality is translated into data, data becomes model behavior, model behavior becomes output, and output becomes evidence for broader claims about knowledge, capability, or reliability. Each layer makes the system easier to use and scale, but each layer also compresses the original conditions into a simpler form.

This is how a local representational error can become systemic. A model output may begin as a limited response generated from compressed representations. After it moves through summaries, recommendations, decisions, dashboards, and user trust, it can harden into a belief about reality. The farther a representation travels from the conditions it was meant to describe, the easier it becomes for coherence to be mistaken for grounding.

Examples Across Systems

Hallucination: A language model may generate information that appears accurate while lacking factual support.

Grounding Failure: A system may have access to relevant information but fail to anchor its output to the evidence provided.

Faithfulness Failure: A model may summarize or transform information in ways that subtly distort the original meaning.

Distribution Shift: A model trained under one set of conditions may encounter environments that differ from the data it learned from.

Benchmark Overfitting: A system may become highly optimized for evaluation metrics while generalization weakens.

Model Collapse: Models trained heavily on synthetic outputs may lose informational diversity and become less connected to original sources.

Conclusion

Hallucination, grounding failure, faithfulness failure, distribution shift, benchmark overfitting, and model collapse are not identical problems. Each has its own technical meaning. But they can also be understood as related forms of representational drift.

The deeper failure is survivable wrongness. The system remains fluent, useful, and operational enough to avoid rejection while its connection to reality weakens.

More coherence does not solve this problem if coherence is no longer constrained by evidence, context, and real-world feedback. The deeper challenge is maintaining fidelity between the representations AI systems use and the realities those representations are supposed to preserve.

Related Phrases and Concepts

- Hallucination
- Grounding
- Faithfulness
- distribution shift
- benchmark overfitting
- alignment failure
- model collapse
- out-of-distribution behavior
- retrieval failure

Semantic Fidelity Project Resources

- Research Library (GitHub)
- Articles & Essays (Substack)
- Semantic Fidelity Glossary
- LLM Failure Modes and Semantic Fidelity