Hallucination, Distribution Shift, Alignment Failure, and Reality Drift

Why AI Systems Continue Working While Gradually Losing Contact With Reality

AI Systems and Reality Drift - Note #3 v1

A. Jacobs - Semantic Fidelity Project

The Basic Pattern

Modern AI systems often appear highly capable. They answer questions, summarize information, write code, produce recommendations, and perform tasks that previously required human judgment.

At first, these systems can appear strongly aligned with their intended purpose. Outputs are useful, benchmarks improve, and users report successful interactions. Over time, however, a familiar pattern can emerge in which systems continue producing coherent outputs while their connection to reality gradually weakens.

Responses may become confidently incorrect. Performance may degrade in unfamiliar situations. Models may remain fluent while becoming less reliable. Systems may continue operating even when their underlying assumptions no longer match the conditions they are supposed to represent.

Different communities describe this challenge using different terminology. It appears in areas such as hallucination, distribution shift, alignment failure, model collapse, concept drift, data drift, and evaluation failure. Although these areas emphasize different failure modes, they often point toward the same structural problem. The system continues functioning while gradually losing alignment with the conditions it was designed to model.

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, and embeddings represent relationships between meaning.

At first, these representations provide useful approximations. The system appears aligned because the representations remain closely connected to the underlying conditions they describe. But as the system is optimized, deployed, reused, and evaluated through those representations, a gap can begin to emerge.

The model may continue producing coherent outputs while becoming increasingly responsive to its own internal patterns rather than the conditions those patterns were meant to capture. The system remains fluent and operational, but fluency is not the same as grounding.

How AI Drift Emerges

Stage 1 — Representation

Training data, objectives, benchmarks, and feedback provide useful representations of the environment. The model learns patterns that remain reasonably connected to reality.

Stage 2 — Optimization

The system becomes increasingly effective at predicting, generating, ranking, or responding. Performance improves, capabilities expand, and the model becomes more useful within the conditions it has learned.

Stage 3 — Decoupling

The representations used by the model begin diverging from the conditions they were originally derived from. New environments appear, user behavior changes, data distributions shift, and evaluations fail to capture emerging weaknesses.

Stage 4 — Drift

The system remains coherent and operational. Outputs continue being generated, and metrics may remain stable. Yet fidelity to reality gradually weakens as the model increasingly reflects its internal optimization environment rather than the conditions it was designed to understand.

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 preserve the surface appearance of understanding while no longer preserving the conditions, relationships, and constraints that made the response meaningful in the first place. This is where AI drift becomes a semantic fidelity problem. The representation remains coherent, but its connection to the underlying reality has degraded.

AI drift also involves constraint collapse. Grounding mechanisms are supposed to constrain model behavior by keeping outputs tied to evidence, context, objectives, or real-world feedback. But when the system becomes increasingly organized around internal representations, those constraints can weaken. The model remains operational, but its outputs lose some of the corrective force that should keep them answerable 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 capability or reliability. Each layer makes the system easier to use and scale, but each layer also compresses the original conditions into a simpler representation.

This is how representational failure can become systemic. A model output may begin as a limited response generated from compressed representations, but after it moves through summaries, recommendations, decisions, and user trust, it can harden into a belief about reality. The further the 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 credible while lacking factual grounding.

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

Alignment Failure: A system may optimize an objective while producing outcomes that diverge from the purpose the objective was meant to serve.

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

Concept Drift: A model may continue using old patterns after the meaning or structure of the environment has changed.

Conclusion

Hallucination, distribution shift, alignment failure, 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.

The challenge is not simply building systems that produce more fluent outputs. More coherence can still drift if it is not constrained by reality. The deeper challenge is maintaining fidelity between the representations AI systems use and the conditions those representations are supposed to preserve.

Related Phrases and Concepts

- Hallucination
- distribution shift
- alignment failure
- model collapse
- concept drift
- data drift
- out-of-distribution behavior
- loss of grounding
- reward hacking
- specification gaming

Semantic Fidelity Project Resources

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