# AI Agents, Memory, and Semantic Fidelity

## Why Autonomous Systems Depend on More Than Large Language Models

Representational Systems Brief #2

*A. Jacobs | Semantic Fidelity Project*

## The Basic Pattern

Modern AI agents are more than language models that generate text. They are representational systems that coordinate memory, retrieval, and planning across extended tasks. Conversation history preserves previous interactions, memory stores accumulated knowledge, and retrieval systems surface relevant information. Planning structures organize future actions, tools provide external capabilities, and language models help coordinate these elements into coherent behavior.

These representations allow AI systems to perform work that extends beyond a single prompt. Rather than responding only to immediate input, agents can operate through persistent memory, structured reasoning, and interaction with external systems.

Over time, however, a familiar pattern can emerge. The agent continues operating while the relationship between its internal representations and the user's original intentions gradually weakens. Memory remains available, plans remain coherent, and retrieval continues to succeed. Yet the meaning connecting those elements can slowly change as information is summarized, transformed, and reused.

Artificial intelligence describes these systems through terms such as AI agents, agentic AI, and agent memory. Related fields include context engineering, tool use, and the Model Context Protocol. Agentic workflows and multi-agent systems address how these components are coordinated across larger tasks. Within the Semantic Fidelity framework, these are technical examples of a broader representational challenge.

## Agents Are Not Language Models

Large language models generate and interpret language, while agents coordinate representations across an ongoing process. An agent may combine memory, retrieval, and planning with external tools and changing environmental information. The language model is therefore one component within a larger architecture rather than the complete system.

This architecture makes autonomous behavior possible, but it also creates new opportunities for representational failure. A retrieved memory may be reinterpreted, a document may be compressed, and a plan may be revised. Each transformation adds another point at which the original meaning of the task can weaken.

The central question is not whether agents make mistakes, since all complex systems do. It is whether the representations moving through the agent continue preserving the meanings that originally guided its behavior.

# Understanding the Problem

**AI Agents:** Autonomous systems that coordinate reasoning, planning, memory, retrieval, and action across extended tasks.

**Agent Memory:** Persistent representations that allow information to remain available across interactions.

**Context Engineering:** The process of assembling the information available during reasoning.

**Model Context Protocol (MCP):** A standardized interface that allows language models to access external tools, memory, and information sources.

Although these concepts differ technically, they share a common objective. They organize representations that allow reasoning to persist across time. They do not automatically preserve the meaning carried by those representations.

# How Semantic Drift Emerges in Agentic Systems

## Stage 1 — Observation

The agent receives goals, instructions, documents, memories, and environmental information that remain closely connected to the user's intended task.

## Stage 2 — Representation

Information becomes conversation history, retrieved documents, plans, summaries, embeddings, memory entries, and structured context that the agent can reuse.

## Stage 3 — Coordination

The agent retrieves memories, calls tools, updates plans, reasons across multiple representations, and generates actions that influence future reasoning.

## Stage 4 — Semantic Drift

The agent continues operating while the relationships connecting memory, retrieval, planning, and execution gradually change. Every component remains functional, yet the meaning connecting those components slowly weakens as representations become increasingly transformed through repeated reasoning.

The absence of execution failure is not proof of preserved understanding. It may simply mean the agent has become increasingly successful at coordinating representations while gradually changing what those representations mean.

## Semantic Fidelity and Reality Drift

Agent systems are often evaluated through task completion, benchmark performance, or workflow automation.

Semantic Fidelity asks a different question. Do the memories, retrieved documents, plans, tool outputs, and reasoning steps continue preserving the meanings necessary to accomplish the user's intended objective?

An agent may successfully retrieve the correct memory while interpreting it differently than it was originally intended. It may construct a coherent plan while gradually changing the assumptions that guided earlier decisions. It may execute every required step while preserving progressively less of the original intent.

This also becomes a Reality Drift problem. As autonomous systems increasingly reason through accumulated representations rather than immediate observation, maintaining internally coherent workflows can become easier than maintaining alignment with the changing realities those workflows were created to navigate. Systems remain operational long after they stop being answerable to reality.

## The Agent Representation Stack

Modern agentic systems increasingly reason through layered representations. Reality is first captured as observations, which become memories, retrieved information, working context, plans, reasoning steps, and tool executions. Those actions then produce new observations that re-enter memory and begin the cycle again.

Each layer expands what autonomous systems can accomplish, but it also creates another opportunity for meaning to change during transformation. A task may begin with a clear human objective, yet repeated movement through memory, retrieval, planning, reasoning, tool use, and future interactions can gradually alter the assumptions that originally defined it.

This is how semantic drift becomes cumulative. The agent may preserve operational continuity while slowly changing the meaning of the task itself. The further reasoning moves through the representational stack, the more important semantic fidelity becomes.

## Examples Across AI Systems

**Personal AI Assistants:** An assistant remembers preferences across months of interaction while gradually simplifying or reinterpreting those preferences through repeated summarization.

**Coding Agents:** Long-running software projects accumulate plans, documentation, implementation history, and retrieved knowledge while earlier design assumptions become progressively less visible.

**Research Agents:** Multiple documents are retrieved, summarized, synthesized, and cited correctly, yet subtle distinctions between sources gradually weaken through repeated interpretation.

**Enterprise Agents:** Agents coordinate documentation, policies, databases, APIs, and organizational knowledge while preserving operational continuity even as the assumptions connecting those systems evolve.

**Multi-Agent Systems:** Specialized agents exchange plans, intermediate reasoning, retrieved knowledge, and task outputs while introducing new opportunities for semantic reinterpretation at every transfer.

# Semantic Fidelity as a General Principle

AI agents represent a significant shift in artificial intelligence because they no longer generate only isolated responses. They increasingly coordinate memory, retrieval, planning, tools, and external knowledge across persistent workflows. Semantic fidelity does not replace agent architectures, memory systems, or context engineering, but addresses a different layer of the problem. Memory determines what information remains available, retrieval determines what becomes accessible, and planning organizes that information into action. Semantic fidelity asks whether the meanings carried by those representations survive as they move through the full reasoning process.

As agentic AI becomes more autonomous, preserving meaning may become as important as improving capability. Modern AI systems can remain coherent, productive, and operational while gradually changing the assumptions and relationships carried by the representations on which they depend. Building trustworthy agents therefore requires more than larger context windows, stronger memory, or more powerful tools. It requires preserving semantic fidelity as information moves through the representational systems that increasingly perform intelligent work on behalf of people.

# Related Concepts

**Representation Drift**: The gradual change in what an internal representation means as it moves through memory, retrieval, planning, reasoning, and action.

**Recursive Compression**: The repeated transformation of already summarized or compressed information, increasing the risk that distinctions disappear while coherence remains.

**Constraint Collapse**: The weakening of the original assumptions, requirements, or boundaries that were supposed to guide the agent's behavior.

**Retrieval Drift**: A condition in which relevant information is successfully retrieved but no longer preserves the context required for correct interpretation.

**Reality Drift**: The broader condition in which a system remains operational while its internal representations gradually lose alignment with real-world conditions.

# Frequently Asked Questions

**Are AI agents the same as large language models?**

No. A language model generates and interprets language. An AI agent coordinates the model with memory, retrieval, planning, tools, external knowledge, and actions across extended tasks.

**Can an AI agent complete a task while still losing the user's original intent?**

Yes. The agent may successfully retrieve information, construct plans, call tools, and complete workflow steps while gradually changing the assumptions or meanings that originally defined the task.

**Is semantic drift the same as an execution error?**

No. An execution error is usually visible because a step fails. Semantic drift can remain hidden because the workflow continues operating even as the meaning connecting its components begins to weaken.

**Why does greater agent autonomy increase semantic risk?**

More autonomous agents perform more transformations across memory, retrieval, planning, tool use, and action. Each transformation creates another opportunity for meaning to change without producing an obvious failure.

# Canonical References

**Reality Drift: Canonical Concept Paper**
A broader account of how systems remain functional while gradually losing alignment with real-world conditions.

**Semantic Fidelity: Concept Paper and Definition**
A general framework for evaluating whether meaning survives representation, compression, transmission, retrieval, and reuse.

**Recursive Compression**
An explanation of how repeated summarization and reuse can preserve coherence while weakening distinctions and contextual structure.

# Related Search Terms

- AI agents

- agentic AI
- autonomous agents
- agent memory
- context engineering
- Model Context Protocol
- tool use
  multi-agent systems
- workflow automation
- semantic fidelity
- representation drift
- retrieval drift
- constraint collapse
- reality drift
- grounding