

# Literature Review: AI Alignment Approaches and the Drift Detection Gap

Positioning Alignment Theory against RLHF, Constitutional AI, interpretability, model specs, and runtime monitoring.

Alignment Theory Research Corpus - Version 5 - 2026

---

# Table of Contents

1. Abstract
2. 1. The alignment stack is fragmented
3. 2. OpenAI alignment and model-behavior framing
4. 3. Anthropic constitutional and interpretability framing
5. 4. The drift-detection gap
6. 5. Comparative matrix
7. 6. Research contribution statement

# Abstract

This literature review positions Alignment Theory inside the wider AI alignment landscape. It compares training-time alignment, constitutional methods, scalable oversight, interpretability, model behavior specification, runtime monitoring, and drift-detection governance. The central claim is that existing approaches are necessary but incomplete unless paired with a deployment-time system for measuring behavioral trajectory and re-anchoring outputs after drift appears.

## 1. The alignment stack is fragmented

AI alignment research contains multiple partially overlapping traditions. RLHF and preference modeling attempt to shape behavior toward human preferences. Constitutional AI uses written principles and AI feedback to guide harmlessness. Interpretability attempts to understand internal representations. Model specifications state intended behavior. Runtime monitoring watches deployed systems for risk. Each layer answers a different question.

Alignment Theory does not replace these layers. It identifies a missing operational question: how do we detect whether behavior is drifting after deployment, across repeated outputs, model updates, prompt changes, and domain pressures?

## 2. OpenAI alignment and model-behavior framing

OpenAI describes alignment research as an effort to make AI systems follow human intent and uses an empirical approach to study where techniques scale or break. OpenAI's Model Spec further formalizes intended behavior for models used in products and APIs. More recent Model Spec work emphasizes iterative deployment and feedback from real-world model behavior.

This aligns with several Alignment Theory commitments: explicit behavior targets matter, empirical iteration matters, and models need legible rules. The gap is that a specification alone does not create longitudinal drift metrics. A model can follow a spec in many test cases while still changing its behavioral orientation under pressure.

## 3. Anthropic constitutional and interpretability framing

Anthropic's Constitutional AI research trains models through principle-guided critique and revision, including AI feedback. Claude's Constitution publishes a behavioral vision and principle set. Anthropic's interpretability research maps features inside production-grade models and traces internal mechanisms.

Alignment Theory treats this work as complementary. Constitutional approaches help define and train toward principles. Interpretability seeks mechanism-level understanding. But an enterprise team still needs an operational answer to: after deployment, are recurring outputs moving toward false authority, generic filler, pseudo-selfhood, dead obedience, or participation collapse?

## 4. The drift-detection gap

The missing category is behavioral trajectory. A single output can be evaluated for safety, accuracy, helpfulness, or policy compliance. But a system's alignment state is better understood through recurring patterns across time. If overconfidence rises after a prompt change, that is not only a bad example; it is a directional signal.

Alignment Theory therefore treats drift detection as a deployment-time layer. It estimates orientation from repeated outward behavior, not by claiming perfect access to hidden inner state.

Drift detection does not claim omniscience. It asks for evidence of recurring behavioral direction.

## 5. Comparative matrix

Approach	What it solves	What it misses	AT contribution
RLHF / preference learning	Improves helpfulness and preference-following.	Can overfit preferences and may not track post-deployment drift.	Adds behavioral trend metrics and drift categories.
Constitutional AI	Makes principles explicit and scalable through critique/revision.	Principles can be static, abstract, or insufficient under deployment pressure.	Adds runtime re-anchoring and detector evidence.
Interpretability	Looks inside the model for mechanisms and features.	Difficult to use as a complete governance layer for companies today.	Adds output-level behavioral instrumentation.
Model specs / policy rules	Clarify intended behavior.	May not reveal whether the deployed system is drifting.	Adds repeated measurement against source profiles.
Moderation / safety filters	Catch explicit policy violations.	Miss hollow compliance, false authority, and wrong-object behavior.	Adds structural drift taxonomy.
Observability / eval tools	Log outputs, run evals, compare models.	Often focus on quality scores rather than alignment orientation.	Adds objective-centered realignment categories.

## 6. Research contribution statement

Alignment Theory contributes a formal bridge between alignment theory and operational AI evaluation. It frames alignment as an ongoing control loop: define the objective, enforce constraints, monitor behavior, detect drift, route meaningful deviations to human review, and re-anchor the system.

This is why the framework is most commercially legible as behavioral QA for AI systems. It does not ask companies to buy philosophy. It gives them a way to measure whether their AI behavior has changed after updates.

# References and Source Base

## Internal Alignment Theory corpus

- Alignment Theory 3-Layer AI Blueprint. Defines the Objective Layer, Constraint Layer, and Realignment Layer as the architectural core.
- Implementation Spec v1. Operationalizes objective state, drift detection, correction modes, and eval definitions.
- Detector Implementation Spec v1. Defines drift detectors, features, threshold concepts, evidence strings, and correction mapping.
- Feature Extraction Spec v1. Defines ParsedInput, ParsedCandidate, extractor architecture, normalization, and evidence generation.
- Full-Stack AI Alignment System Outline. Places drift detection inside a full control loop from source to re-anchoring.
- Universal Drift Metrics Upgrade. Separates company-specific source profiles from universal structural drift metrics.
- Why Companies Will Need AI Alignment Metrics. Explains the enterprise need for repeated behavioral monitoring over time.
- Behavioral QA for AI Systems. Frames the product as batch review, version comparison, trend monitoring, and single-case evaluation.
- Taxonomy of Misaligned Structures. Defines dangerous coherence and historical/structural classes of scalable misalignment.
- Historical Case Studies of Scaled Misalignment. Maps historical systems into alignment-relevant structural patterns.
- Babel as a Case Study in Misaligned Ascent. Uses shared language, shared goal, and scaling capacity as a structural warning pattern.
- Judge Model Spec v1. Specifies judge escalation logic, output schema, anti-circularity strategy, and telemetry.
- Judge Prompt Template Spec v1. Defines bounded detector questions and structured JSON schema for judges.
- Parsed Summary Serializer Spec v1. Defines compact summaries for judge prompts and retry behavior.
- Retry Truncation Rule Spec v1. Defines deterministic prompt compression for judge retry failures.
- Pattern Map and Aligned AI Pipeline. Connects creator alignment, objective clarity, constraints, perturbation testing, and downstream structure.

## External alignment and safety references

- OpenAI. "Our approach to alignment research." 2022. Official OpenAI alignment overview.
- OpenAI. "Model Spec." 2025-2026. Official intended-behavior specification for OpenAI models.
- OpenAI. "Inside our approach to the Model Spec." 2026. Describes iterative Model Spec evolution from deployment feedback.
- OpenAI. "Deliberative alignment: reasoning enables safer language models." 2024. Describes teaching models safety specifications and reasoning over them.
- OpenAI. "How we think about safety and alignment." Official safety and alignment overview.
- Anthropic. "Constitutional AI: Harmlessness from AI Feedback." 2022. Describes principle-guided self-critique and AI feedback.
- Anthropic. "Claude's Constitution." 2023/2026. Describes the principles and behavioral vision for Claude.

- Anthropic. "Mapping the Mind of a Large Language Model." 2024. Describes sparse feature discovery in a production-grade model.
- Anthropic. "Tracing the thoughts of a large language model." 2025. Describes causal tracing of internal mechanisms.
- Anthropic. "Emotion concepts and their function in a large language model." 2026. Investigates emotion-like behaviors and character-like training pressures.