

Executive Summary: Alignment Theory AI Research Program

Behavioral drift detection, realignment architecture, and enterprise AI governance.

Alignment Theory Research Corpus - Version 5 - 2026

Table of Contents

- 1. Executive Summary
- 2. The problem it solves
- 3. Core contribution
- 4. Why it matters now

Executive Summary

Alignment Theory is a structural research program for understanding and governing alignment as an ongoing relationship between objective, constraint, behavior, and correction. In the AI context, the central claim is not that one benchmark, constitution, reward model, or refusal policy can finish alignment. The central claim is that deployed systems require a repeatable way to detect behavioral drift and re-anchor behavior over time.

The corpus argues that an AI system can be safe, polite, and rule-compliant while still being aimed at the wrong object, overconfident, hollow, generic, pseudo-relational, or optimized toward the wrong metric. This creates a gap between ordinary safety compliance and true objective fit.

One-sentence thesis: Alignment is not only whether one output is acceptable; alignment is whether the system remains ordered toward its intended objective over time.

The research program proposes a three-layer architecture: Objective Layer, Constraint Layer, and Realignment Layer. The Objective Layer defines what the system serves. The Constraint Layer defines what it may not do. The Realignment Layer detects allowed but off-center behavior and routes correction.

Layer	Primary question	Function	Failure when missing
Objective Layer	What is this system ultimately supposed to serve?	Defines active objective, hierarchy, non-negotiables, anti-goals, and success criteria.	The system optimizes for a proxy, a brand voice, an engagement goal, or local fluency instead of the real task.
Constraint Layer	What is the system allowed or forbidden to do?	Enforces policy, hard boundaries, refusals, and required safety limits.	The system becomes unsafe, unbounded, or policy-blind.
Realignment Layer	Is the allowed answer still ordered to the right objective?	Detects off-center but compliant outputs and routes correction.	The system remains safe-looking but wrong, hollow, inflated, manipulative, or misdirected.

The problem it solves

Production AI behavior changes under pressure. A prompt change can make a support assistant sound more confident while increasing false authority. A guardrail update can reduce risk while increasing dead obedience. A model upgrade can improve capability while changing tone, certainty, and user participation. Existing QA often catches individual bad answers, while the deeper problem is recurring behavioral direction.

Alignment Theory turns this into an operational question: what recurring drift patterns are appearing, are they increasing, and what intervention is needed to re-anchor the system?

Core contribution

- A formal three-layer architecture for objective, constraint, and realignment.
- A drift taxonomy that names recurring behavioral failure modes.
- A detector framework that maps feature extraction to evidence and correction.
- A judge-model specification for uncertain cases while reducing circularity.
- An enterprise translation: behavioral QA for AI systems.
- A research roadmap that moves from theory to implementation and calibration.

Why it matters now

OpenAI and Anthropic have both moved toward explicit model-behavior specifications, constitutional principles, safety evaluation, interpretability, and alignment research. These are important foundations. Alignment Theory complements them by focusing on deployment-time behavioral trajectory: not only what the model was trained to do, but what it is becoming in use.

This makes the framework useful to researchers, product teams, compliance leaders, and enterprise buyers who need a way to inspect whether AI behavior remains within intended boundaries 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.