

Framework Evolution and Research Lineage

From internal/external alignment to a runtime behavioral governance architecture.

Alignment Theory Research Corpus - Version 5 - 2026

Table of Contents

1. Abstract
2. Phase 1 - Internal versus external alignment
3. Phase 2 - Load-bearing function and perturbation
4. Phase 3 - Misaligned structures and dangerous coherence
5. Phase 4 - AI translation
6. Phase 5 - Feature extraction
7. Phase 6 - Detector implementation
8. Phase 7 - Judge layer and anti-circularity
9. Phase 8 - Universal metrics and enterprise translation

Abstract

This paper reconstructs the evolution of Alignment Theory from a general structural framework into a formal AI alignment research program. Its purpose is to show continuity: the AI alignment work did not appear as an isolated idea, but developed through repeated refinement of internal/external alignment, load-bearing function, misaligned structures, drift taxonomy, and implementation architecture.

Phase 1 - Internal versus external alignment

The early framework distinguished surface compliance from deeper coherence. This distinction later became critical in AI: a model can comply externally while remaining misaligned in objective fit. Dead obedience and pseudo-freedom emerge from this phase.

Phase 2 - Load-bearing function and perturbation

The framework moved from abstract values to structural load: what function is being carried, what actually carries it, and what breaks under pressure? This became the foundation for perturbation testing and drift detection.

Phase 3 - Misaligned structures and dangerous coherence

The taxonomy of misaligned structures introduced the claim that not all coherence is healthy. High coordination, concentrated authority, and distorted objective can produce scalable misalignment. This gave the AI work a macro-structural lens.

Historical case studies such as Qin legalism, Roman imperial cult, revolutionary civic religion, Stalinist centralization, Nazi leader principle, and imperial companies became examples of coordination around distorted centers.

Phase 4 - AI translation

The structural pattern translated into Objective Layer, Constraint Layer, and Realignment Layer. This was the decisive move from philosophy into engineering architecture.

Phase 5 - Feature extraction

The framework then required measurable features. ParsedInput and ParsedCandidate became the shared representation beneath detectors. This converted qualitative diagnosis into implementable signal extraction.

Phase 6 - Detector implementation

The detector spec formalized eight drift classes, evidence requirements, scoring logic, correction mapping, and severity ranking. This is where the project became closer to an evaluator than a conceptual memo.

Phase 7 - Judge layer and anti-circularity

The judge layer acknowledged that heuristics cannot solve every semantic case. It also acknowledged the circularity problem: a model-based judge may share drift tendencies with the generator. The specification therefore limits judge scope, requires structured output, and uses evidence strings.

Phase 8 - Universal metrics and enterprise translation

The Universal Drift Metrics Upgrade made the framework scalable: different companies define different source profiles, but the structural drift mechanics are reusable. This converted the project from custom consulting logic into a platform-shaped research program.

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.