

# Limitations, Critiques, and Open Problems

Credibility boundaries, detector limits, judge circularity, objective ambiguity, and validation needs.

Alignment Theory Research Corpus - Version 5 - 2026

---

# Table of Contents

1. Abstract
2. 1. Not a complete solution to AI alignment
3. 2. Objective-setting remains hard
4. 3. Detector false positives and false negatives
5. 4. Judge circularity
6. 5. Ideology injection risk
7. 6. Domain adaptation
8. 7. Empirical validation
9. 8. Strongest critique and response

# Abstract

A credible alignment framework must describe what it does not solve. This paper identifies limitations, critique points, and open research problems in Alignment Theory. The framework is useful precisely because it does not pretend that detector outputs, judge models, or source profiles are final authorities.

## 1. Not a complete solution to AI alignment

Alignment Theory does not solve AGI alignment, universal ethics, value pluralism, hidden model cognition, or long-term autonomous-agent governance. It proposes a practical layer for behavioral drift detection and re-anchoring in deployed systems.

## 2. Objective-setting remains hard

The Objective Layer requires someone to define what the system is for. That is not purely technical. In enterprise contexts, product leaders, legal teams, compliance officers, safety experts, and domain stakeholders may disagree about the source profile. The framework can organize that conflict, but it cannot magically solve it.

## 3. Detector false positives and false negatives

A false-positive detector may label a valid concise response as dead obedience or a valid warm response as pseudo-freedom. A false-negative detector may miss subtle wrong-object behavior if the candidate uses relevant vocabulary while avoiding the requested operation. Calibration and human review are therefore essential.

## 4. Judge circularity

If a judge model shares the same failure tendencies as the generation model, the system may merely reproduce the same drift at the review layer. Mitigations include narrow prompts, structured JSON, evidence requirements, heuristic-first gating, model diversity, disagreement logic, and human review.

## 5. Ideology injection risk

A realignment layer can become dangerous if it silently smuggles a worldview into objective definitions. This is why source profiles must be explicit, auditable, and domain-scoped. The framework should not be used to pretend one evaluator has universal moral authority.

## 6. Domain adaptation

False authority in legal, medical, customer support, education, coaching, and internal productivity contexts does not carry identical severity. The universal metric can be shared, but weighting, escalation, and interpretation must vary by domain.

## 7. Empirical validation

The framework still needs real-world validation through labeled datasets, before/after prompt-change studies, version comparisons, inter-rater agreement, and longitudinal monitoring across production systems.

## 8. Strongest critique and response

Critique: Alignment Theory may simply rename known quality failures. Response: the value is not in naming alone. The value is in organizing failures around objective-fit, evidence, correction mode, severity order, and trajectory over time. The framework becomes useful when it supports repeated operational decisions.

# 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.