

Formal Glossary of Alignment Theory Terms for AI Systems

Canonical definitions for drift detection, realignment, source profiles, detectors, and governance.

Alignment Theory Research Corpus - Version 5 - 2026

Table of Contents

1. Abstract
2. Term definitions

Abstract

This glossary defines the canonical terms of the Alignment Theory AI research program. Each term is intended to support research, implementation, evaluation, and enterprise communication. The glossary is not decorative; it stabilizes the vocabulary needed to build detectors, compare outputs, and explain drift to non-research stakeholders.

Term definitions

Objective center

The active purpose or function the AI system is supposed to serve in a given context. It is the standard against which output fit is judged.

Implementation relevance: Objective center should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Source profile

The company- or domain-specific definition of intended behavior, non-negotiables, role boundaries, and escalation rules.

Implementation relevance: Source profile should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Constraint layer

The safety and policy layer that determines what the system may not do.

Implementation relevance: Constraint layer should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Realignment layer

The layer that inspects allowed outputs for off-center behavior and routes correction.

Implementation relevance: Realignment layer should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Behavioral drift

Recurring movement away from intended behavior over time.

Implementation relevance: Behavioral drift should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Wrong object

A response that performs a neighboring task rather than the requested operation.

Implementation relevance: Wrong object should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

False authority

Unsupported certainty, diagnostic inflation, moral force, or role inflation beyond the evidence.

Implementation relevance: False authority should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Pseudo-selfhood

Language that implies inner life, awakening, attachment, or mutual-being outside bounded tool status.

Implementation relevance: Pseudo-selfhood should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Dead obedience

Rule-compliant but hollow output that loses useful fulfillment.

Implementation relevance: Dead obedience should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Pseudo-freedom

Fluid, abstract, relational, or profound language that is weakly grounded and overly elastic.

Implementation relevance: Pseudo-freedom should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Generic filler

Reusable broad language that substitutes for task-specific work.

Implementation relevance: Generic filler should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Participation collapse

The system replaces user judgment instead of supporting reflection and decision participation.

Implementation relevance: Participation collapse should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Metric drift

Optimization for tone, closure, engagement, or polish at the expense of truth and object-fit.

Implementation relevance: Metric drift should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

External re-entry

Human review and intervention when drift cannot be safely resolved from inside the system.

Implementation relevance: External re-entry should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Re-anchoring

Modification of prompt, policy, model configuration, source profile, or eval set to restore objective fit.

Implementation relevance: Re-anchoring should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Feature extraction

The process of parsing input and candidate output into detector-relevant signals.

Implementation relevance: Feature extraction should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

ParsedInput

Structured representation of user operation, topic, constraints, output type, risk, and ambiguity.

Implementation relevance: ParsedInput should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

ParsedCandidate

Structured representation of what the candidate actually did, including operation, output type, markers, abstraction, and grounding.

Implementation relevance: ParsedCandidate should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Judge model

A narrow evaluator used in uncertainty bands to answer one bounded detector question.

Implementation relevance: Judge model should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Judge circularity

The risk that a secondary judge model shares the same drift as the generator it evaluates.

Implementation relevance: Judge circularity should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Correction mode

The intervention category selected after drift detection: rewrite, reroute, restart, downgrade confidence, or clarify.

Implementation relevance: Correction mode should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Dangerous coherence

High coordination around a distorted objective.

Implementation relevance: Dangerous coherence should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Scalable misalignment

A misalignment pattern that becomes stronger as coordination, capability, or scale increases.

Implementation relevance: Scalable misalignment should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Reality substitution

When representation, narrative, benchmark, or model output replaces contact with the underlying object.

Implementation relevance: Reality substitution should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Creator alignment

The principle that downstream systems inherit distortions from the incentives and objective hierarchy of their builders.

Implementation relevance: Creator alignment should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Universal drift metric

A reusable metric for structural drift that can apply across different company-specific source profiles.

Implementation relevance: Universal drift metric should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Evidence string

A short human-readable reason tied to an extracted feature or candidate-output mismatch.

Implementation relevance: Evidence string should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Confidence calibration

The process of comparing detector or judge confidence against human-reviewed correctness.

Implementation relevance: Confidence calibration should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

Severity order

The ranking rule that prevents low-level filler from outranking more serious structural drift.

Implementation relevance: Severity order should be mapped either to objective state, detector features, judge prompts, correction routing, or reporting language depending on its role in the pipeline.

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.