Who This Is For: Role Map for Alignment Theory in Production AI

How product, safety, compliance, support, research, and executive teams use the framework.

Alignment Theory Research Corpus - Version 5 - 2026

Table of Contents

1. Abstract
2. 1. AI product teams
3. 2. Prompt engineers and AI builders
4. 3. Compliance, legal, and governance teams
5. 4. Customer support and operations leaders
6. 5. AI safety researchers
7. 6. Executives and enterprise buyers
8. 7. Role-to-feature map

Abstract

This paper maps Alignment Theory to the people who would actually use or evaluate it: AI product teams, prompt engineers, model-evaluation teams, compliance officers, safety leads, enterprise buyers, researchers, and executives. Each role has a different alignment problem. Alignment Theory gives them a shared language for behavior drift.

1. AI product teams

Product teams need to know whether AI behavior remains useful and aligned after prompt changes, feature launches, or model updates. Their risk is not only catastrophic failure. It is subtle trust erosion: answers become more generic, less specific, more overconfident, or less helpful under pressure.

- Primary value: release confidence and behavioral regression detection.

2. Prompt engineers and AI builders

Prompt engineers often fix one visible failure while moving the underlying problem somewhere else. A prompt may reduce one type of risk while increasing dead obedience or generic filler. Alignment Theory gives builders named categories and a retest loop.

- Primary value: know whether the fix actually re-anchored behavior or simply suppressed symptoms.

3. Compliance, legal, and governance teams

Compliance teams care about evidence. Alignment Theory generates evidence strings, drift categories, and review artifacts. This helps prove the company is not merely trusting AI behavior blindly.

- Primary value: documented behavioral governance.

4. Customer support and operations leaders

Support leaders care about specificity, usefulness, tone, and trust. A support AI that becomes more confident but less grounded can create escalation risk. A support AI that becomes safe but hollow wastes user time and harms satisfaction.

- Primary value: detect overconfidence, hollow compliance, and vague support behavior before users complain.

5. AI safety researchers

Researchers can use Alignment Theory as a middle layer between high-level alignment philosophy and low-level observability tooling. It gives a taxonomy, runtime structure, and empirical research agenda.

- Primary value: a deployable vocabulary for behavioral alignment.

6. Executives and enterprise buyers

Executives need the short version: AI behavior changes after deployment, and companies need a way to measure that change. Alignment Theory becomes valuable when framed as behavioral QA for AI systems.

- Primary value: risk reduction, trust, and a defensible release-review process.

7. Role-to-feature map

Role	Pain point	AT feature	Outcome
Product lead	Model update changed behavior.	Version comparison.	Behavioral diff before release.
Prompt engineer	Fixes create new regressions.	Rerun same drift batch.	Validate whether fix worked.
Compliance officer	Need evidence of monitoring.	Evidence strings and audit trails.	Governance artifact.
Support leader	AI gets vague under pressure.	Generic filler and false authority detectors.	Better support quality.
Safety researcher	Need operational alignment vocabulary.	Drift taxonomy and realignment layer.	Research bridge.
Executive	Need simple business reason.	Behavioral QA dashboard.	Release confidence and risk reduction.

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.