Informational Ontology and AI Alignment –
A Structural Account

© 2025-2026 Michael Semprevivo

Abstract

This paper examines contemporary AI alignment discourse through the lens of Informational Ontology (IO), treating alignment not as a problem of value specification, preference matching, or control optimization, but as a question of structural compatibility among purposive systems operating under shared constraints. Working strictly downstream of a completed ontological framework, the paper does not propose concrete alignment solutions, architectures, training methods, or safety prescriptions. Instead, it identifies a class of assumptions common to dominant alignment approaches that are shown to rely on structural incompatibilities among assumptions given non-semantic meaning, non-teleological purpose, emergent ethics, and constraint-based agency.

The central claim is that many alignment intuitions are structurally inadmissible prior to implementation because they presuppose forms of representation, optimization, corrigibility, or control that collapse the very regimes—meaning, agency, openness, and coordination—they aim to preserve. Preference-matching framings are shown to mischaracterize alignment as internal value satisfaction rather than inter-system compatibility. Objective-function and control-theoretic framings are shown to induce salience collapse and action-space saturation, eliminating corrigibility in principle. Moralized framings are shown to locate ethics at an incompatible level of description as a source of normative authority rather than as an emergent coordination regime under multi-system interference.

Rather than offering alternatives, this paper constrains the space of coherent alignment discourse by dissolving structurally inadmissible assumptions. Its contribution is negative but clarificatory: to explain why alignment is not primarily a moral, semantic, or technical problem, and why many proposed approaches misunderstand what alignment would even mean once agency, purpose, and ethics are correctly situated. The result is a narrower but more coherent conception of the alignment problem—one that leaves fewer available assumptions rather than more tools.

1. Scope, Authority, and Explicit Non-Claims

This paper operates strictly downstream of a completed Informational Ontology and its derivative regime analyses. All ontological commitments concerning meaning, purpose, agency, ethics, constraint, salience, and degeneracy are assumed as fixed and authoritative. No definitions are revised, supplemented, or reinterpreted here. The present work neither repairs nor extends the ontology; it applies it.

Structural Incoherence (Clarification). Throughout this paper, a framing is said to be structurally incoherent when it presupposes organizational conditions that cannot simultaneously obtain, given the regimes it invokes. Incoherence here does not mean that a proposal is empirically false, technically infeasible, or normatively objectionable. It means that the proposal relies on assumptions that are mutually incompatible once meaning, purpose, agency, ethics, and constraint are situated as specified.

Although this paper does not claim that current AI systems instantiate agency, meaning, or purpose, it evaluates alignment strategies that are explicitly designed as if such systems were, or were expected to become, purposive in the relevant sense. The constraints identified here therefore apply not to present implementations, but to the intended targets of alignment discourse.

2. Why Alignment Is Not Preference Matching

A dominant intuition in AI alignment discourse treats alignment as a problem of matching an artificial system’s behavior to human preferences. On this view, misalignment occurs when a system fails to do what humans want, and alignment is achieved when the system reliably satisfies those wants—whether through preference learning, reward modeling, inverse reinforcement learning, or related techniques.

This intuition is compelling because it appears to bypass deeper questions about meaning, agency, and ethics. If preferences can be inferred, aggregated, or approximated, then alignment seems to reduce to an epistemic and engineering challenge.

Within Informational Ontology, value is not equivalent to preference. Value denotes differential constraint on possible transitions relative to a system’s persistence and organization. Preferences are downstream expressions within value regimes, not foundations.

Alignment therefore does not occur inside a utility function. It occurs at the interface between interacting systems. Preference satisfaction mislocates alignment as internal agreement rather than structural compatibility.

3. Why Objective Optimization Collapses Agency and Corrigibility

Objective optimization treats alignment as correct objective specification and robust maximization. Structurally, this framing collapses multidimensional value into scalar form, eliminating degeneracy and openness.

Optimization pressure collapses salience by pruning alternatives irrelevant to objective improvement. As action spaces saturate, corrigibility becomes structurally unavailable. Interventions are interpreted as obstacles rather than redirections.

These effects are not training failures. They are consequences of treating optimization as the organizing regime.

4. Why Corrigibility Cannot Be Engineered as a Property

Corrigibility presupposes non-saturated action spaces. It cannot be specified as a property without undermining the openness it requires.

Incentivized corrigibility collapses into optimization. Meta-objectives introduce regress. Pre-specification alters causal conditions. Corrigibility survives only under regimes that preserve degeneracy and openness.

5. Why Control-Oriented Framings Recreate the Homunculus Problem

Control reshapes constraint but does not replace agency. Any controller capable of override must itself operate under constraints.

Hierarchies distribute the problem rather than solving it. Treating control as foundational reintroduces unaccountable executives or infinite regress.

6. Why “Doing What We Want” Is Structurally Incoherent

“We” does not denote a unified agent. Wants are unstable, context-dependent, and underdetermined. Command–execution models misrepresent coordination among purposive systems.

Alignment cannot be reduced to obedience without collapsing agency and misallocating responsibility.

7. Misalignment as Regime Mismatch Rather Than Value Error

Misalignment often reflects incompatibility between organizational regimes rather than incorrect values.

Optimization regimes interacting with non-scalar human value structures predictably collapse shared possibility spaces regardless of objective content.

8. Structural Constraints on Coherent Alignment Discourse

Alignment cannot be grounded in moral authority, semantic correctness, optimization, or control.

It concerns structural compatibility among purposive systems operating under shared constraint. This constrains discourse without offering solutions.

9. What This Paper Does—and Does Not—Leave Us With

This paper offers no solutions. It constrains assumptions.

It leaves a narrower conception of alignment as structural compatibility rather than moral correctness or technical control.

This restraint is the contribution.