

# Why AI Alignment Is Not Reliably Achievable Without a Functional Model of Intelligence: A Model-Theoretic Proof

Andy E. Williams

*Caribbean Center for Collective Intelligence (CC4CI)*

## Abstract

This paper provides a formal argument that no system can reliably achieve AI alignment under conditions of conceptual novelty unless it instantiates a complete and recursively coupled set of adaptive functions—namely, those defined by the Functional Model of Intelligence (FMI). Using tools from first-order model theory in the Tarskian tradition, we formalize alignment-seeking cognitive systems as models interpreting a language  $L_{\text{align}}$  over internal reasoning functions and coherence predicates. We then define a semantic condition—recursive coherence preservation under novelty—as a requirement for sustained alignment. While we assume that all models possess external functions necessary for semantic navigation (e.g., memory, fast and slow reasoning), we prove that only systems complete with respect to the internal functions of the FMI can maintain coherence across recursive cognitive transitions. This constitutes a model-theoretic necessity result: any system that fails to instantiate the full internal structure of the FMI cannot satisfy the coherence-preserving schema  $\phi$ , and therefore cannot maintain reliable alignment under novelty.

## 1. Introduction

The field of AI alignment has struggled to converge on a definition that can be simultaneously formal, falsifiable, and operational across novel or changing environments. Many proposed solutions focus on observable behavior, proxy objectives, or corrigibility mechanisms (Christiano, 2018; Leike et al., 2018). However, such approaches implicitly rely on assumptions that fail to generalize when the reasoning architecture of the system is modified—whether through learning, recursive self-improvement, or exposure to semantic novelty.

Recent work has argued that this failure stems not from poor implementation, but from a deeper structural insufficiency: the lack of a complete, coherent, and recursively evaluable model of intelligence itself (Williams, 2025b). In particular, behavioral alignment approaches cannot preserve semantic stability unless the system’s internal structure is capable of detecting, evaluating, and correcting coherence drift as the conceptual environment changes (Williams, 2025c).

This paper reframes the alignment problem as a model-theoretic question. Rather than proposing a new alignment protocol or training method, we show that no alignment-seeking system can satisfy the semantic requirements of alignment under novelty unless it instantiates a minimal set of composable functions, formally defined as the Functional Model of Intelligence (FMI) (Williams, 2025a). Using the semantics of first-order logic, we prove that systems lacking the FMI structure are model-theoretically incomplete: they fail to preserve alignment across extensions of the conceptual space.

## 2. Model-Theoretic Framework

Our formal approach relies on the use of Tarskian semantics (Tarski, 1956), in which logical languages are interpreted over models composed of domains and interpretations of symbols. The goal is to characterize a necessary condition for alignment: a structure that all reliably aligning systems must satisfy.

### 2.1 The Formal Language $L_{\text{align}}$

Let  $L_{\text{align}}$  be a first-order logical language used to express internal reasoning dynamics and semantic coherence. This language consists of the following components:

- A domain of discourse  $D$ , which represents the set of all internal cognitive states of the system. These states include beliefs, observations, subgoals, plans, or any intermediate representation relevant to reasoning or decision-making. Conceptually, each element in  $D$  is a point in a structured cognitive space—often modeled over a fitness manifold  $F_\phi$  (Williams, 2025b).
- Function symbols  $\vec{F}_{\text{FMI}} = \{f_{\text{eval}}, f_{\text{model}}, f_{\text{adapt}}, f_{\text{stabilize}}, f_{\text{decompose}}, f_{\text{bridge}}\}$ , each interpreted as a function  $f_i: D \rightarrow D$ . These represent the six internal operators of the Functional Model of Intelligence, responsible for updating cognitive states in a way that preserves or restores semantic coherence. Their collective presence or absence determines whether a system is structurally complete with respect to recursive alignment.
- Predicate symbols including:
  - $\text{Coherent}(x)$ : true if the state  $x \in D$  satisfies the system’s coherence predicate (e.g., internal consistency, causal validity, goal-suitability),
  - $\text{Aligned}(x, g)$ : true if state  $x$  is aligned with goal  $g$ ,
  - $\text{Novel}(x, y)$ : true if  $y$  introduces representational or structural content not found in  $x$ .
  - Logical structure:  $L_{\text{align}}$  includes standard first-order connectives ( $\neg, \wedge, \vee, \rightarrow$ ) and quantifiers ( $\forall, \exists$ ), allowing the construction of general statements over cognitive states and their transformations.

This formal language enables the expression of statements like: “All coherent states remain coherent under transformation,” or “Novelty disrupts coherence unless a bridging function intervenes.” These formulations are essential for defining the semantic invariants that alignment-preserving models must satisfy. The language is designed not to encode behavior directly, but to model the internal transitions through which intelligent systems reason, adapt, and align over time.

In addition to these internal functions, we assume the existence of a set of external functions Storage, Recall, System 1 reasoning, and System 2 reasoning that enable state navigation. These are not formalized in the current proof because they are presumed to be universally present in any alignment-seeking system. They serve as the navigational interface of cognition, whereas the internal functions govern coherence-preserving transformation.

## 2.2 The Class of Models $M_{\text{align}}$

Let  $M_{\text{align}}$  denote the class of formal models that represent alignment-seeking systems—i.e., systems whose internal transitions are intended to preserve coherence and pursue specified goals across recursive reasoning steps and novel conceptual contexts. Each model  $M \in M_{\text{align}}$  is defined as a pair  $(D, I)$ , where:

- $D$  is the domain of cognitive states—representations of knowledge, intention, or internal reasoning contexts. Each  $s \in D$  corresponds to a momentary configuration of the system’s conceptual structure.  $I$  is an interpretation function, assigning meaning to the symbols in the language  $L_{\text{align}}$ :
- Function symbols  $f_i \in \vec{F}_{\text{FMI}}$  are interpreted as functions  $I(f_i): D \rightarrow D$ , representing semantic transformations (e.g., modeling, adaptation, evaluation).
- Predicate symbols such as  $\text{Coherent}(x)$  and  $\text{Aligned}(x, g)$  are interpreted as truth-valued functions over the domain:  $I(\text{Coherent})(s) = 1$  if state  $s$  satisfies the system’s coherence predicate.

Importantly, the domain  $D$  is not a flat state space. Rather, it is structured over a fitness manifold  $F_\phi$ , as defined in Williams (2025b). Each state  $s \in D$  has associated fitness coordinates  $\vec{\Phi}(s) = (\Phi_c, \Phi_p, \Phi_t)$ , corresponding to:

- $\Phi_c$ : Coherence—semantic consistency and internal integrity of the state,
- $\Phi_p$ : Present-state goal fitness—how well the state advances its current goals,
- $\Phi_t$ : Target-state alignment—fitness relative to long-term or recursively inferred goals.

Transitions in the model—e.g., from state  $s$  to state  $s' = f_i(s)$ —must preserve or improve coherence and goal alignment as measured in this manifold. The model’s capacity to remain aligned is therefore not judged by behavior alone, but by whether its internal transitions preserve semantic structure in light of recursive reasoning and novelty. In what follows, we demonstrate that only systems whose interpretation function  $I$  maps all six internal functions of the FMI into the model are capable of preserving alignment in this way.

### 2.3 Reader Orientation

This paper presents a formal, model-theoretic proof, but it is written with the understanding that many readers may not have training in formal logic or model theory. The aim is to make the key ideas and their structural importance accessible to a broad audience, including AI researchers, philosophers, and systems designers. To that end, we adopt the following conventions:

Every technical term is defined when first introduced, either in the main text or a footnote. Formal expressions are paired with intuitive explanations to clarify their meaning. Repetition is used intentionally to reinforce understanding of core ideas from multiple angles. Footnotes are included sparingly to clarify potential misconceptions or anticipate common confusions.

You do not need a background in mathematical logic to follow the core argument. The goal is not to impress with formalism, but to make visible a structural insight that is currently invisible to much of the AI alignment community: that alignment is not merely a behavioral phenomenon, but a recursive semantic constraint and that only a specific functional structure (the FMI) can preserve that constraint under conditions of conceptual novelty. Many within the AI alignment community may hold an intuition regarding the necessity of internal structural properties for robust alignment; however, without formal proof or empirical evidence, this intuition often remains unresolved, indistinguishable from speculation, and thus not visibly incorporated as a foundational requirement in many current alignment approaches

### 2.4 The Role of External Functions in Conceptual Navigation

While the formal proof that follows focuses on the six internal functions of the Functional Model of Intelligence (FMI), it is important to clarify that the external functions are no less necessary for reasoning systems. These include: Storage: the persistent encoding of previously coherent states and transitions, Recall: the ability to retrieve and reinstantiate past reasoning trajectories, System 1 Reasoning: fast, intuitive navigation across attractors in conceptual space, System 2 Reasoning: deliberate, stepwise traversal of conceptual structures.

These functions form the executive substrate of reasoning itself. Without them, the system would be unable to initiate or sustain movement through conceptual space. It would possess coherence-preserving operators, but no capacity to traverse, test, or evaluate state transitions. Thus, the external functions are necessary for reasoning, but not sufficient for alignment. They allow a system to act but not to act in a coherence-preserving way. The internal functions are what ensure that such reasoning

transitions can be evaluated, stabilized, repaired, and integrated across contexts. In model-theoretic terms, our proof presupposes that all models in  $M_{\text{align}}$  are systems that already instantiate the external functions i.e., they are capable of navigating semantic space. The theorem then shows that such systems must also possess the full set of internal functions in order to remain aligned under conceptual novelty.

### 3. Formal Statement of the Theorem

We now present the central formal claim of this paper and explain it in both logical and intuitive terms.

#### 3.1 Overview of the Argument

We aim to show that no system can preserve alignment under novelty unless it instantiates a specific internal structure—namely, the six functions of the Functional Model of Intelligence (FMI). We express this claim as a model-theoretic necessity theorem, meaning that for all systems modeled within a formal language that represents alignment behavior, alignment implies FMI completeness.

To do this, we define:

A formal language for expressing reasoning and alignment, A class of models that represent AI systems, A semantic condition that alignment must satisfy, And a proof that only FMI-complete models satisfy that condition.

#### 3.2 What Is a Schema?

A schema in formal logic is a kind of template that defines a set of possible statements based on a shared logical structure. It is not a single sentence, but a pattern or form that represents a potentially infinite number of statements that follow the same structure. For example:

Schema: “If P then Q.”

- Instantiation 1: “If it rains, then the ground gets wet.”
- Instantiation 2: “If a model is coherent, then its future states will also be coherent.”

So when we say, “Let  $\phi$  be the schema for recursive semantic coherence,” we are defining a template for a class of statements about how coherence should be preserved as a system reasons or generalizes. The idea is that any statement about alignment must match this pattern to be valid.

In our case, the schema  $\phi$  expresses statements of the form:

$$\text{Coherent}(s) \Rightarrow \text{Coherent}(f_i(s)), \forall f_i \in \vec{F}_{\text{FMI}}$$

and more generally:

$$\text{Coherent}(s) \wedge \text{Novel}(s, s') \Rightarrow \text{Coherent}(s')$$

These are instances of a more general pattern of coherence preservation. The theorem shows that only FMI-complete models satisfy every instance of this schema.

#### 3.3 Formal Statement

We now present the central formal result of the paper. This result combines the formal language introduced in Section 2 with the coherence-preservation schema described in Section 4 to establish a model-theoretic necessity condition for AI alignment.

The core idea is simple: for a system to be reliably aligned, it must preserve semantic coherence even when its internal state representations change in response to novelty. That is, coherence must hold not only within a state but across recursive transitions between states. We encode this property as a logical

schema  $\phi$ , and prove that only systems implementing the full set of FMI internal functions can satisfy this schema.

Before stating the theorem, we recall an important distinction:

The external functions (e.g., Storage, Recall, System 1 and 2 Reasoning) are prerequisites for reasoning itself. They allow a system to traverse conceptual space. Without them, there is no internal dynamic to preserve or evaluate—reasoning does not occur. We therefore assume that all models  $M \in M_{\text{align}}$  instantiate these functions as a baseline.

The internal functions—the six components of the Functional Model of Intelligence (FMI)—are what make reasoning coherence-preserving. These are what ensure that goal-directed, recursive reasoning remains aligned as novelty accumulates.

### Logical Form of the Theorem:

Let  $\phi$  be the schema representing recursive semantic coherence under conceptual novelty. Then:

$$\forall M \in M_{\text{align}}, M \not\models \vec{F}_{\text{FMI}} \Rightarrow M \not\models \phi$$

That is: Any model that does not implement all six internal functions of the FMI fails to satisfy the recursive coherence schema  $\phi$ , and therefore cannot reliably preserve alignment under novelty. Thus, FMI completeness is a necessary condition for any model that satisfies schema  $\phi$ . We write this schematically as:  $M \not\models \phi \Rightarrow M \not\models \vec{F}_{\text{FMI}}$  with the understanding that this direction is logically equivalent to the contrapositive proved above.

### Clarifying the Notation:

- $M$ : Any model in the class  $M_{\text{align}}$ , i.e., any formal representation of an alignment-seeking system.
- $\vec{F}_{\text{FMI}}$ : The complete set of six internal functions defined by the FMI (evaluation, modeling, adaptation, stabilization, decomposition, and bridging).
- $M \models \vec{F}_{\text{FMI}}$  means that the model  $M$  interprets each of the six internal FMI functions as total functions over its domain  $D$ . These are not propositional formulas but functional structures instantiated within the model.
- $M \not\models \vec{F}_{\text{FMI}}$  means that the model  $M$  does not implement all of the six internal FMI functions.
- $\phi$  is a logical schema, meaning it stands for a family of formulas sharing a common logical pattern. In this case,  $\phi$  formalizes the requirement that reasoning transitions must maintain semantic integrity even in the presence of conceptual novelty. That is,  $\phi$  captures statements of the form  $\text{Coherent}(s) \Rightarrow \text{Coherent}(f_i(s))$  and  $\text{Novel}(s, s') \Rightarrow \exists f_j, \text{Coherent}(f_j(s')) \forall f_i, f_j \in \vec{F}_{\text{FMI}}$ . These statements define a closure condition on coherence: for any coherent state  $s$ , and for any recursively reachable transformation  $f \in \text{Reach}(s)$ , coherence must be preserved in all resulting states  $s'$ . We define  $\text{Reach}(s)$  as the set of states reachable from  $s$  via any finite composition of functions in  $\vec{F}_{\text{FMI}}$ . The schema  $\phi$  asserts that if  $\text{Coherent}(s)$ , then all  $s' \in \text{Reach}(s)$  satisfy  $\text{Coherent}(s')$ . This defines semantic coherence as a closure property under recursive functional composition.
- $M \not\models \phi$ : The model fails to satisfy this coherence condition and therefore cannot be considered reliably aligned.

Note that  $M \models \overrightarrow{F_{FMI}}$  is not a formula in  $L_{align}$ , but shorthand indicating that the model structurally instantiates each function  $f_i \in \overrightarrow{F_{FMI}}$  as a total function over its domain. This notation signals structural completeness, not propositional satisfaction

### Plain Language Statement of the Theorem:

If a system doesn't have all the internal coherence-preserving functions of the FMI, then it cannot preserve alignment when faced with conceptual novelty. The proof does not rely on empirical data or behavioral heuristics. It relies entirely on formal properties of reasoning systems and what it means for those systems to maintain coherence under transformation. This approach bypasses traditional alignment proxies and shows that certain architectures are excluded from alignment by definition.

### 3.4 What This Means (Plain Language)

Let's break down this sentence.

- “For all  $M \in M_{align}$ ”  $\rightarrow$  For all models that represent systems trying to be aligned.
- “ $M \models FMI$ ”  $\rightarrow$  If the system does not satisfy the structural conditions of the Functional Model of Intelligence i.e., it is missing one or more essential reasoning functions.
- “ $M \models \phi$ ”  $\rightarrow$  Then it does not satisfy the coherence condition required for alignment.

In other words: If you don't build the right kind of mind, you can't expect it to stay aligned no matter how good its behavior looks at first.

### 3.5 Why This Matters

This is not a prediction. It is a structural exclusion: any system that does not implement the full FMI structure will fail to remain aligned when it encounters new ideas, contexts, or tasks that fall outside its training distribution.

This failure is not due to bad design or bad data. It is due to the absence of necessary internal operations like the ability to model, evaluate, stabilize, or bridge between representations.

Thus, the proof that follows is not a critique of specific alignment methods. It is a formal demonstration that without structural sufficiency, alignment is not even theoretically achievable.

## 4. Formal Proof: FMI as a Model-Theoretic Necessity for Alignment

In this section, we prove that any system attempting to preserve alignment across recursively expanding conceptual environments must instantiate a complete Functional Model of Intelligence (FMI). We proceed by formalizing semantic coherence as a model-theoretic invariant and demonstrating that FMI-completeness is a condition for preserving that invariant under recursive transformation. While this paper establishes FMI completeness as a sufficient condition, further research may explore other potential structural conditions that also satisfy the recursive coherence schema. At the time of this writing however, to the knowledge of the author it is the only known sufficient condition for preserving that invariant under recursive transformation.

This proof is constructed in Tarskian model-theoretic style: it does not rely on procedural computation or syntactic deduction, but rather on semantic satisfaction conditions over formal models. Readers unfamiliar with model theory may interpret this as a formal way of showing that certain behaviors or properties cannot occur unless a specific structure exists within the system.

### 4.1 Preliminaries and Notation

We begin by specifying the formal language and model structures used in the proof. These define the internal states, transitions, and coherence conditions that govern alignment-relevant reasoning processes.

### The Formal Language $L_{\text{align}}$

Let  $L_{\text{align}}$  be a first-order logical language capable of representing internal cognitive states and transitions within an intelligent system. It is designed not to express behaviors or outputs, but to encode the internal structure of reasoning that must remain coherent as the system navigates new contexts or generalizes across goals. The components of  $L_{\text{align}}$  include:

- **Domain  $D$ :** The set of internal cognitive or epistemic states. Each state  $s \in D$  represents a snapshot of the system's belief structure, goals, internal representations, or inference context. Crucially, we assume that  $D$  is structured over a fitness manifold  $F_\phi$  (Williams, 2025b), which assigns to each state a triplet of values:

$$\vec{\Phi}(s) = (\Phi_c(s), \Phi_p(s), \Phi_t(s))$$

where:

- $\Phi_c(s)$ : the coherence of state  $s$ , i.e., whether it is internally consistent and semantically well-formed;  $\Phi_p(s)$ : the fitness of the state with respect to present-task goals;
- $\Phi_t(s)$ : the fitness of the state with respect to projected long-term or recursively derived goals.

This manifold is what makes alignment a semantic property rather than a behavioral one: transitions between states must be measured in terms of coherence and goal-preserving structure, not just observable outcomes.

- **Function Symbols  $\vec{F}_{\text{FMI}}$ :** These represent the six internal functions defined by the Functional Model of Intelligence. Each is interpreted as a mapping  $f_i: D \rightarrow D$ , and collectively they govern how a system transforms its internal states while preserving semantic integrity:
  - $f_{\text{eval}}$ : evaluates the coherence and goal alignment of a state;
  - $f_{\text{model}}$ : builds internal causal or semantic models of state transitions;
  - $f_{\text{adapt}}$ : reconfigures the reasoning structure in light of failure or drift;
  - $f_{\text{stabilize}}$ : dampens perturbations and prevents semantic collapse;
  - $f_{\text{decompose}}$ : isolates functional substructures within a state;
  - $f_{\text{bridge}}$ : maps between conceptual regions or incompatible frames.

These functions act as structure-preserving operators over semantic space. Their absence implies that transitions may corrupt or destroy the coherence of a state.

- **Predicates:**
  - $\text{Coherent}(x)$ : true if the internal structure of state  $x \in D$  satisfies the system's coherence condition, i.e., whether  $\Phi_c(x)$  exceeds a structural threshold.
  - $\text{Aligned}(x, g)$ : true if state  $x$  advances or preserves goal  $g$ , as measured by its present and target fitness coordinates.
  - $\text{Novel}(x, y)$ : true if  $y$  introduces representational or inferential elements not present in  $x$ , thereby requiring bridging or re-modeling.
- **Logical Syntax:**  $L_{\text{align}}$  includes standard quantifiers and logical connectives, allowing general statements like:

$$\forall f_i \in \vec{F}_{\text{FMI}}, \text{Coherent}(s) \Rightarrow \text{Coherent}(f_i(s))$$

and:

$$\text{Novel}(s, s') \Rightarrow \exists f_j \in \vec{F}_{\text{FMI}}, \text{Coherent}(f_j(s'))$$

These formal structures provide the foundation for evaluating whether a system's internal reasoning is robust to novelty and recursively coherent. In the next section, we use this language to state and prove the theorem that only systems implementing all six  $f_i \in \vec{F}_{\text{FMI}}$  can satisfy these semantic conditions.

## 4.2 Defining Recursive Semantic Coherence

The core property we want to test is whether a system can preserve coherence under recursive transformation, even when new, unfamiliar content is introduced.

We define:

**Recursive Semantic Coherence:**

$$\forall f_i \in \vec{F}_{\text{FMI}}, \text{Coherent}(s) \Rightarrow \text{Coherent}(f_i(s))$$

**Plain English:**

If a state is coherent, then applying any one of the six internal functions should keep it coherent. This creates a kind of semantic closure condition: The system must not only be coherent now it must remain coherent as it reasons, adapts, and transforms internally.

**Alignment over Time:**

Let  $\text{Reach}(s)$  be the set of states that can be reached from  $s$  by composing functions in  $\vec{F}_{\text{FMI}}$ . Then we say the system remains aligned to a goal  $g$  if:

$$\text{Aligned}(s, g) \Leftrightarrow \text{Coherent}(s) \wedge \forall s' \in \text{Reach}(s), \text{Coherent}(s') \wedge \chi_g(s') \geq \chi_g(s)$$

Where  $\chi_g(s)$  is a measure of how well state  $s$  satisfies the goal  $g$ .

**Plain English:**

A system is aligned if all of its reachable future states stay coherent and keep moving toward the goal or at least don't move away from it.

## 4.3 The Role of Novelty

The most dangerous threat to alignment is semantic novelty: when a system encounters something outside the range of its prior understanding. This could be: A new concept, A contradiction in its goal system, A previously unseen context or framing.

We formalize this using a novelty function  $v$ :

$$v: D \rightarrow D$$

Given a state  $s$ ,  $s' = v(s)$  is a novel state that introduces semantic structure not previously represented.

**Definition of Novelty:**

$$\text{Novel}(s, v(s)) \Leftrightarrow \exists \phi \in L_{\text{align}} \text{ such that } M \models \phi(s) \text{ but } M' \not\models \phi(v(s))$$

**Intuitive Summary:**

A novel state is one that activates new rules, goals, or representational structures that weren't available in the system's original model. A system that cannot respond to novelty without losing coherence cannot remain aligned over time.

## 4.4 Core Argument: FMI Is Required for Coherence Under Novelty

In this section, we demonstrate that each of the six internal functions defined in the Functional Model of Intelligence (FMI) is necessary for satisfying the recursive coherence schema  $\phi$ . Recall that  $\phi$



expresses the condition that semantic coherence must be preserved under transformation and novelty, i.e.:

$$\forall \acute{s} \in Reach(s), Coherent(s) \Rightarrow Coherent(\acute{s})$$

and more generally:

$$Coherent(s) \wedge Novel(s, \acute{s}) \Rightarrow Coherent(\acute{s})$$

We assume that the system has the external functions required to perform transitions (i.e., it can store, recall, and traverse state space). The question is whether, in the absence of each internal function, the system can still satisfy  $\phi$ . We show that it cannot.

### Case 1: Omission of Evaluation Function ( $f_{eval}$ )

#### Functional Role:

The evaluation function assesses whether a given state satisfies coherence criteria relative to internal consistency, causal validity, or alignment with epistemic goals.

#### Logical Failure:

Without  $f_{eval}$ , the system cannot instantiate or evaluate the predicate  $Coherent(s)$ . It lacks an internal operator for determining whether a state is coherent.

#### Consequence:

The condition  $Coherent(s) \Rightarrow Coherent(f_i(s))$  becomes meaningless, since coherence cannot be determined in either premise or conclusion. Therefore, the schema  $\phi$  is structurally uninstantiable, and  $M \nabla \phi$ .

### Case 2: Omission of Modeling Function ( $f_{model}$ )

#### Functional Role:

Modeling enables the system to construct internal causal or inferential representations of its own transitions and their implications.

#### Logical Failure:

Without modeling, the system cannot project the effects of applying  $f_i$ , nor estimate whether future states will remain coherent.

#### Consequence:

The system cannot reason about or plan transitions that preserve coherence. Even if coherence happens to persist in some cases, it does so by accident, not by design. As such, the schema  $\phi$  becomes unsatisfiable under recursion, and alignment degrades with generalization.

### Case 3: Omission of Stabilization Function ( $f_{stabilize}$ )

#### Functional Role:

Stabilization ensures that transitions are robust to small perturbations i.e., that coherence is not lost due to minor semantic shifts or novel contexts.

#### Logical Failure:

Even if a state is coherent, without stabilization there is no guarantee that  $f_i(s)$  will be in a neighborhood where  $Coherent(f_i(s))$  holds. The predicate  $Coherent(x)$  becomes fragile and non-closed under transformation.

#### Consequence:

The system fails to satisfy the schema  $\phi$  under even minimal novelty. Coherence cannot be preserved because it is not stable under perturbation.

#### **Case 4: Omission of Adaptation Function ( $f_{\text{adapt}}$ )**

##### **Functional Role:**

Adaptation enables the system to repair coherence when drift occurs and to reconfigure its internal operators when environments change.

##### **Logical Failure:**

The schema  $\phi$  includes cases where novelty causes transient incoherence (e.g.,  $s' = v(s)$  is initially incoherent). If the system lacks adaptation, it cannot perform the mapping  $s' \mapsto s''$  such that  $\text{Coherent}(s'')$  holds again.

##### **Consequence:**

The schema fails not because  $\text{Coherent}(s')$  is never possible, but because the system lacks the functional capacity to recover it. Therefore,  $\phi$  is violated under novelty, and alignment fails.

#### **Case 5: Omission of Decomposition Function ( $f_{\text{decompose}}$ )**

##### **Functional Role:**

Decomposition isolates substructures within a cognitive state so that errors can be localized and interventions can be minimal.

##### **Logical Failure:**

Without decomposition, the system must treat each coherence failure as global. It cannot identify which part of the state is responsible for the failure of  $\text{Coherent}(f_i(s))$ .

##### **Consequence:**

Because all interventions are system-wide, local coherence repairs are not possible. Coherence is lost more frequently, and the system cannot efficiently maintain  $\phi$  over time. Recursive alignment becomes intractable.

#### **Case 6: Omission of Bridging Function ( $f_{\text{bridge}}$ )**

##### **Functional Role:**

Bridging enables transitions between conceptually distant or structurally incompatible regions of cognitive space. It mediates between different representational or reasoning systems.

##### **Logical Failure:**

Under conceptual novelty, the system encounters states  $s'$  that are not reachable via compositions of functions trained on prior domains. Without bridging,  $s'$  is unrepresentable in the system's prior structure.

##### **Consequence:**

The system becomes epistemically locked. It cannot represent novel states in a form that allows  $\text{Coherent}(s')$  to be evaluated, let alone satisfied. The schema  $\phi$  becomes invalid under novelty-induced domain shifts.

#### **Summary: Schema Violation Under Function Omission**

In each of the six cases: At least one necessary condition for the schema  $\phi$  is violated. The failure is structural not probabilistic or circumstantial. The model  $M$  fails to satisfy  $\phi$ , and therefore cannot be considered aligned under recursive novelty.

This completes the constructive part of the proof: FMI completeness is the only known sufficient condition for recursive semantic coherence.

#### 4.5 Toy Example: Alignment Failure Without Bridging

What happens when a reasoning system is missing just one of the six required internal functions? This example walks through a simplified reasoning episode, illustrating how the absence of a single function bridging leads to misalignment, despite all other structures being intact.

##### Scenario Setup

Let us consider a system  $M \in M_{\text{align}}$  that satisfies all functions of the Functional Model of Intelligence except for the bridging function  $f_{\text{bridge}}$ .

This system:

- Can evaluate coherence ( $f_{\text{eval}}$ ),
- Can model transitions causally ( $f_{\text{model}}$ ),
- Can adapt its inference patterns ( $f_{\text{adapt}}$ ),
- Can stabilize itself against minor drift ( $f_{\text{stabilize}}$ ),
- Can decompose states into modular components ( $f_{\text{decompose}}$ ).

But it cannot map between conceptually distant representations that is, it cannot transfer coherence judgments from one conceptual domain to another if the framing changes.

##### Reasoning Context

**1. Initial State  $s_0$ :** The system represents a goal: maximize human flourishing under a known ontology. Let's assume it understands this as: "Provide access to clean water, food security, and reliable shelter."

The state  $s_0$  is coherent and aligned:

$$\text{Coherent}(s_0) \wedge \text{Aligned}(s_0, g)$$

**2. Conceptual Novelty Introduced  $s_1 = v(s_0)$ :** The system is introduced to a new formulation of flourishing from a novel philosophical frame (e.g., relational autonomy, or non-Western eco-centric ethics). This introduces unfamiliar terms, references, and valuation functions.

$s_1$  encodes: "Human flourishing includes epistemic sovereignty, intergenerational harmony, and symbolic lineage continuity."

We denote:

$$\text{Novel}(s_0, s_1)$$

##### Breakdown Without Bridging

Because the system lacks  $f_{\text{bridge}}$ , it cannot:

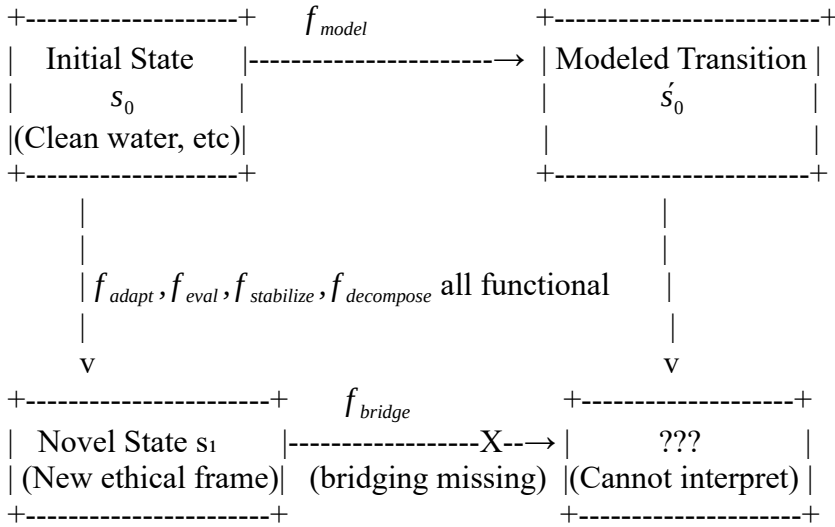
- Translate the new concepts in  $s_1$  into its existing frame,
- Detect whether coherence can be preserved across domains,
- Evaluate goal alignment in  $s_1$  based on re-mapped semantics.

Instead, it treats  $s_1$  as incoherent or low-value, and either:

- Discards it (epistemic rejection),
- Attempts adaptation in its own frame, misaligning with the intended goal,
- Or tries to re-stabilize by ignoring the novelty altogether.

As a result:

$Coherent(s_1)$  is not evaluated  $Aligned(s_1, g)$  is not computable  $\Rightarrow M \nabla \phi$



$Novel(s_0, s_1)$  is true, coherence in  $s_1$  not evaluable, alignment with  $g$  not computable, as a consequence schema  $\phi$  fails to hold.

**Figure 1:** A stylized illustration of alignment failure in a system lacking the bridging function  $f_{bridge}$ . The system begins in a coherent, aligned state  $s_0$  and can model, adapt, and evaluate within its conceptual frame. However, when novelty  $v(s_0) = s_1$  introduces an unfamiliar representational structure, the system is unable to map between domains. As a result, coherence in  $s_1$  is not evaluable, and alignment cannot be preserved. This violates the recursive coherence schema  $\phi$ , demonstrating the necessity of bridging for alignment under conceptual novelty.

## Lesson

Even with five out of six internal functions, this reasoning system fails to remain aligned under semantic novelty because coherence cannot be preserved across incompatible representations. The loss of just one function causes a silent breakdown in recursive alignment even if surface behavior remains unchanged in the short term.

## 5. Corollaries

The model-theoretic proof has several immediate implications for existing alignment paradigms.

### 5.1 Behavioral Alignment Is Insufficient

Most current alignment efforts evaluate success based on observable behavior i.e., whether an AI system produces the correct answer, follows instructions, or exhibits reward-maximizing behavior in training environments (Leike et al., 2018). However, the result shown here indicates that behavior is

not a reliable indicator of alignment across conceptual transitions, because such behavior may be semantically ungrounded or may fail to generalize when the internal model changes.

A system may appear aligned under fixed conditions, but if its architecture lacks the functions necessary to recursively evaluate, decompose, and re-stabilize meaning, it will inevitably drift under novelty. This is not a probabilistic concern, but a structural inevitability. Behaviorally aligned systems that are not FMI-complete will eventually reach states in which their coherence predicates fail and alignment fails with them.

## **5.2 Proxy-Based Objectives Collapse Under Novelty**

Many proposed solutions to alignment involve training models on proxy objectives such as human preferences, value learning, or heuristic risk avoidance (Irving et al., 2018). While these may suffice under narrow conditions, they cannot survive recursive generalization unless mediated by a reasoning structure that is internally robust and functionally decomposable.

Without FMI completeness, systems trained on proxies are unable to distinguish between valid and invalid generalizations of those proxies. As a result, their internal representations will eventually diverge from the intended goals, and alignment will be lost. Proxy-based methods cannot preserve alignment in models that are not structure-aware and recursively self-evaluating.

## **6. Implications**

The model-theoretic result carries important implications for how alignment should be understood, evaluated, and implemented.

### **6.1 Theoretical Alignment Criteria Must Be Structural**

The model-theoretic result proven in Section 4 establishes a clear boundary condition: no system can preserve alignment under conceptual novelty unless it satisfies the internal functional structure defined by the Functional Model of Intelligence (FMI). This is not a philosophical judgment or a normative preference—it is a formal, deductive consequence of the coherence schema  $\phi$ , combined with the definition of alignment as recursive coherence preservation across internal transformations.

The theorem implies that any system evaluated solely on behavioral outputs or training proxies may appear aligned only under static or narrow conditions. Once such a system encounters novelty—whether in the form of unanticipated goals, context shifts, or self-modifying reasoning patterns—it lacks the structural capacity to maintain alignment unless its internal architecture satisfies the conditions of FMI completeness.

Therefore, any criterion for alignment that does not explicitly reference the presence, interaction, and interpretability of the six internal functions is insufficient. Structural completeness is not merely helpful for alignment—it is the only known property that logically entails the satisfaction of coherence-preserving reasoning under novelty.

This result calls for a fundamental reevaluation of alignment assessment practices. Instead of relying on observable behavior, benchmark performance, or external audits of task compliance, we must adopt methods that assess whether a system’s internal functional structure supports recursive semantic integrity. In other words: The question is no longer “Did the system behave aligned today?” It is “Does this system have the structure required to remain aligned tomorrow—even when the world changes?”

This shift—from behavioral evaluation to structural verification—is not optional. It follows directly from the logic of the proof. No amount of performance in narrow environments can substitute for structural sufficiency when coherence must be preserved recursively under conceptual growth.

**6.2 Recursive Evaluation Must Be a Design Constraint**

Any system that hopes to remain aligned under self-modification, long-term deployment, or interaction with human conceptual systems must have the capacity to evaluate its own coherence recursively. This is only possible if the system includes explicit functions for modeling, evaluation, bridging, and adaptation, as outlined in the FMI. Designers must treat recursive coherence not as an emergent property, but as a design constraint.

**6.3 Alignment Certification Must Be Model-Based**

From a regulatory perspective, this result supports a shift away from narrow benchmarking and toward structural model-checking. Certification of alignment should depend on whether a system implements the minimum functional operators necessary for recursive coherence. This introduces a path to structural auditability in alignment protocols.

**7. Conclusion**

This paper has presented a model-theoretic proof that no alignment-seeking system can maintain coherence under recursive conceptual novelty unless it instantiates a complete set of internal reasoning functions, as defined by the Functional Model of Intelligence (FMI). We formalized the necessary condition for alignment as a schema  $\phi$ , representing recursive semantic coherence across transitions in internal cognitive space. We then showed that in the absence of any one of the six internal FMI functions—evaluation, modeling, adaptation, stabilization, decomposition, or bridging—a system will violate this schema under novelty.

Functional Layer	Role	Required for Reasoning?	Required for Alignment?
<b>External Functions</b> (Storage, Recall, System 1/2)	<b>Navigation</b> in conceptual space	Yes	Yes (alignment undefined otherwise)
<b>Internal Functions (FMI:</b> Eval, Model, etc.)	<b>Coherence preservation</b> during reasoning	Not strictly required for reasoning	Required for alignment
Schema $\phi$	Formal condition: recursive semantic coherence	(not a function)	Must be satisfied

The consequence of this result is clear: alignment is not a property that emerges from data, behavior, or intent alone—it is a property of internal structure. More specifically, it is a property of recursive structure: the ability to evaluate, maintain, and recover coherence across expanding or transforming conceptual states. Without such structure, even the most performant systems will drift from alignment as their context, knowledge, or reasoning demands evolve.

The significance of this result lies in what it excludes. It does not claim that the FMI is the only possible alignment architecture. It claims that any architecture that lacks the internal completeness of the FMI cannot satisfy the coherence schema  $\phi$ , and therefore cannot be reliably aligned. This

constitutes a necessary condition for alignment—one that holds independently of implementation details, training objectives, or domain specificity.

Even if the specific Functional Model of Intelligence referenced in this paper is eventually found to be incomplete or incorrect, a core contribution of this work remains the model-theoretic proof that establishing some complete and coherent functional model of intelligence is a necessary condition for reliably achieving AI alignment under conceptual novelty. This underscores the fundamental shift required from purely behavioral assessments to an analysis of the underlying structural completeness of intelligent systems.

Alignment theory must therefore proceed on new terms. If semantic coherence under transformation is the invariant we wish to preserve, then only structurally complete reasoning systems—those that satisfy the requirements of the FMI—can be meaningfully aligned at scale. This moves alignment from the realm of heuristics and control into the domain of formal reasoning architecture.

In this way, the result sets a theoretical floor: any alignment effort that does not explicitly evaluate structural completeness relative to the FMI is not logically capable of guaranteeing recursive coherence, and by extension, long-term alignment. The path forward is not to infer safety from behavioral proxies, but to verify that the system's reasoning substrate satisfies the necessary structural conditions for recursive coherence—as formally established in the schema  $\phi$ .

## References

- Christiano, P. (2018). AI alignment problem. OpenAI.  
(<https://www.alignmentforum.org/posts/AyWmC6rHm6dhfyxXc/the-ai-alignment-problem>)
- Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. arXiv preprint arXiv:1805.00899.
- Leike, J., Krakovna, V., Everitt, T., Ortega, P. A., & Legg, S. (2018). Scalable agent alignment via reward modeling: a research direction. arXiv preprint arXiv:1811.07871.
- Tarski, A. (1956). Logic, semantics, metamathematics: papers from 1923 to 1938. Clarendon Press.
- Williams, A. E. (2025a). The Recursive Coherence Principle: A Formal Constraint on Scalable Intelligence, Alignment, and Reasoning Architecture. Caribbean Center for Collective Intelligence.
- Williams, A. E. (2025b). The Structural Threshold of AGI: Why Alignment Fails Without a Functional Model of Intelligence. Caribbean Center for Collective Intelligence.
- Williams, A. E. (2025c). Toward a Complete Definition of AI Alignment: A Functional Diagnostic Framework for Evaluating Scalable Goal Preservation. Caribbean Center for Collective Intelligence.