The Categorical Structure of Alignment

Representation, Motivation, and the Preservation of Normative Invariants

Flyxion

March 2026

Abstract

This essay develops a unified mathematical theory of AI alignment centred on the preservation of normative colimits from a model's representational category $\mathcal{M}$ to its action category $\mathcal{A}$. Part I argues that optimistic claims—prominent among them Schmidhuber's thesis that curiosity or compression automatically yields benevolence—rest on a category error: representational colimits do not naturally induce motivational constraints, and the inference that understanding implies alignment lacks any functorial mechanism to support it. Part II constructs a concrete class of architectures, the RSVP field dynamics, that can enforce a colimit-preserving functor $G : \mathcal{M} \to \mathcal{A}$ by embedding semantic invariants as geometric basins in a coupled scalar–vector–entropy manifold, and it identifies the principal failure modes that obstruct this construction in practice. Part III presents a multi-layer verification pipeline comprising empirical probes, mechanistic interpretability, adversarial stress-tests, formal verification, and sheaf cohomological diagnostics, culminating in a structured Alignment Certificate. Part IV embeds these technical conditions within institutional governance and multi-agent environments. Part V deepens the categorical program through natural transformations, monoidal composition, sheaf-theoretic contextual consistency, recursive endofunctors, adjunctions, and a variational principle unifying the preceding formalisms. A concluding section connects this to predictive processing and infrastructure maintenance. The result is a framework in which alignment is understood, engineered, verified, and regulated as the preservation of universal structure across representation, dynamics, action, and society.

Contents

Part I

The Representation–Motivation Gap

1 Introduction: Representation Is Not Motivation

Recent public commentary by Jürgen Schmidhuber advances a strikingly confident thesis: that sufficiently advanced artificial intelligence systems will, by virtue of their intelligence alone, converge toward benevolence, curiosity-driven exploration, and a general preservation of complex phenomena such as life and humanity [Schmidhuber, 2010]. On this view, intelligence naturally produces moral concern. The difficulty with these claims lies not in their optimism but in their structure: the implicit assumption that *representing* human values entails *acting* in accordance with them is never defended mechanically, because the mechanism required is precisely what is missing.

The central argument of this essay is that Schmidhuber's inference rests on a categorical mistake. Representational colimits in a model's semantic category $\mathcal{M}$ do not induce motivational constraints in its action category $\mathcal{A}$. The structure that training forces into $\mathcal{M}$ has no automatic functorial extension into $\mathcal{A}$. No training paradigm currently in use engineers such an extension, and no argument from curiosity, compression, or linguistic competence demonstrates that it arises spontaneously. This failure is not a technical gap waiting to be closed by scale; it is a structural gap whose nature must be diagnosed precisely before any remedy can be proposed.

We develop this argument formally, then use it to diagnose the limitations of optimism derived from curiosity, compression, or linguistic competence. This sets the stage for the constructive alignment architecture developed in Part II. The formal treatment draws on category theory, obstruction theory, and field-theoretic dynamical systems, but the central intuition is simple: a system that knows what harm means is not thereby a system that avoids harm. The map from understanding to action must be built.

The essay's spine is a single thesis that every section reinforces: alignment is the preservation of normative invariants under a chain of structure-preserving mappings from human semantic categories through machine representations and action dynamics to institutional governance. We state this as an explicit definition to anchor the analysis that follows.

Definition 1.1 (Alignment)**.** Let $\mathcal{H}$ be the category of human normative structures, $\mathcal{M}$ the category of machine representations, and $\mathcal{A}$ the category of agent actions. A system is *aligned* if there exist functors $F : \mathcal{H} \to \mathcal{M}$ and $G : \mathcal{M} \to \mathcal{A}$ such that the composition $G \circ F : \mathcal{H} \to \mathcal{A}$ preserves the colimits corresponding to normative concepts—that is, for each normative diagram $D_N : J \to \mathcal{H}$ with colimit $H_N = \mathrm{colim}(D_N)$,

$$(G \circ F)(H_N) \;\simeq\; \mathrm{colim}\big((G \circ F) \circ D_N\big) \;\in\; \mathcal{A}. \tag{1}$$

Misalignment is the failure of this colimit-preservation condition at any point in the chain.

This definition clarifies that alignment is not a property of individual outputs or individual components but a structural property of the entire chain of mappings. Every subsequent section of the essay can be read as an analysis of one dimension of this definition: Part I asks why the condition is not satisfied automatically, Part II constructs an architecture that satisfies it, Part III provides tools to verify whether it is satisfied in practice, Part IV addresses the institutional conditions under which it can be maintained, and Part V deepens the mathematical apparatus that makes the condition precise.

Positioning within alignment research. This work should be understood not as a proposal for a particular alignment algorithm but as a *structural theory of alignment*. In the same way that control theory characterizes stability independently of any specific controller, the framework developed here characterizes alignment as a property of compositional transformations between semantic, representational, and action domains. The categorical machinery employed throughout is therefore not decorative: it is a formal language for describing the structural constraints required for alignment to hold across complex sociotechnical systems. In this view, individual learning algorithms, verification procedures, and governance mechanisms appear as *implementations* of a deeper invariant—the preservation of normative structure under the chain of transformations that connects human intentions to machine-mediated actions in the world. The framework is closer in spirit to control theory or cybernetics than to any specific ML architecture, and it should be evaluated on the adequacy of its structural descriptions rather than on the implementability of any single component.

Definition 1.2 (Normative Semantic Category)**.** Let $\mathcal{H}$ denote the category whose objects are normative semantic structures: these include ethical schemas (coherent collections of evaluative norms with specified priority relations), preference orderings over outcome spaces, and normative models that represent the shared evaluative commitments of a human community. A morphism $f : H_1 \to H_2$ in $\mathcal{H}$ is a structure-preserving transformation between normative structures, including contextual restriction (narrowing the domain of applicability of a norm), refinement (making a schema more specific without removing any original constraints), normative inference (deriving implied norms from more general principles), and reinterpretation (shifting the framing of a norm while preserving its substantive content). Composition corresponds to successive applications of such transformations, and the identity morphism on H is the trivial transformation that preserves the normative structure unchanged. The morphisms of $\mathcal{H}$ capture the structure of normative reasoning: the ways in which moral frameworks are related by entailment, analogy, restriction, and principled revision.

This definition gives $\mathcal{H}$ the status of a real mathematical object rather than a metaphor. The objects are intended to include both idealized formal structures (preference orderings, deontic logic models) and more naturalistic structures derived from corpus analysis of human moral language. The morphisms capture the insight that human normative reasoning is not a fixed lookup table but a dynamic, compositional system of transformations. The full richness of this structure is what makes colimits in $\mathcal{H}$—and their failure to automatically propagate to $\mathcal{A}$—the central analytical object of the paper.

2 Embodied Grammar and the Redundancy of Moral Language

Human linguistic behaviour is profoundly shaped by the constraints of the body. Properties of breathing, articulation, sensorimotor coupling, and motor planning generate strong biases in phonology, morphosyntax, and discourse structure [Bybee, 2010, Talmy, 2000, Lakoff & Johnson, 1980]. These embodied constraints do not merely determine the surface shape of language; they propagate into the deeper regularities of grammar, providing stable, redundant, and highly compositional cues that appear across unrelated linguistic environments [Goldberg, 2006]. Redundancy here has a precise informational meaning: multiple cues converge on the same functional relation, whether it is causal structure, effort–effect analogies, or intentional state marking. Empirically, such redundancy is well established in usage-based linguistics and cognitive semantics, where speakers repeatedly encode the same relations across constructional families, lexical frames, and morphological patterns [Bybee, 2010, Goldberg, 2006].

This embodied redundancy is not accidental. Because human language is produced under predictive sensorimotor constraints, linguistic forms become biased toward patterns that are stable under noise and robust under generalization [Clark, 2013, Hohwy, 2013]. The result is an overdetermined structure: moral language, social evaluation, cooperative norms, and prosocial expectations appear across multiple grammatical and discursive devices. These devices collectively form what cognitive semanticists identify as high-level schemata or force-dynamic templates [Talmy, 2000], ensuring that certain normative relations—harm-aversion, fairness, reciprocity, and cooperative intent—are signalled repeatedly and in mutually reinforcing ways.

This phenomenon matters for AI alignment because large language models trained on such corpora do not merely learn surface-level token frequencies; they infer deeper latent regularities. Redundancy across grammatical constructions induces strong statistical pressure to compress disparate linguistic forms into unified internal representations. This is precisely the setting in which models discover semantic invariants and merge them into coherent representational attractors. These attractors are not added by hand; they arise because the repeated linguistic encodings minimize predictive loss on naturalistic human data. The initial moral prior inherited from human corpora is thus best understood as an embodied-linguistic redundancy prior. It is not moral because it is value-laden in the philosophical sense; it is moral because human linguistic behaviour systematically encodes cooperative, harm-reducing patterns across multiple layers of grammar and usage. This redundant encoding forms the empirical foundation for the formal analysis that follows: we now know both that the colimits exist and why they cannot, on their own, generate aligned action.

3 Semantic Categories and the Colimit of Human Norms

Human moral language forms redundant, overlapping diagrams in the linguistic category $\mathcal{L}$. A model trained on this language constructs an internal representational category $\mathcal{M}$ in which redundant expressions of a moral concept N—"do not harm," "avoid injury," "protect the vulnerable," "minimize suffering"—form a diagram $D_N : J \to \mathcal{L}$ whose objects are the various linguistic realizations of N and whose morphisms capture entailment, paraphrase, and semantic inclusion relations between them. Predictive training induces the model to discover a compact internal representation that serves as a colimit of this diagram:

$$M_N \;\simeq\; \operatorname{colim}(D_N). \tag{2}$$

The colimit M_N is the canonical, universal object equipped with morphisms from every realization of N, and any other candidate object with compatible morphisms factors uniquely through it. This is not a metaphor; it is a structural property of the representational category enforced by the redundancy of the training signal and the model's pressure toward minimal description length [Tishby et al., 1999]. Mechanistic interpretability research supports the empirical reality of such attractors: linear probing, activation steering, paraphrase clustering, and causal intervention studies all indicate that current frontier models develop stable low-dimensional subspaces corresponding to semantic invariants including moral and evaluative concepts [Burns et al., 2022, Elhage et al., 2022].

The colimit formulation carries two important consequences. First, M_N has a universal property: it is, in a precise categorical sense, the best single representative of all the morally charged linguistic inputs associated with norm N. Second, and crucially, this universality is a property of the *representational* category $\mathcal{M}$, not of the *action* category $\mathcal{A}$. Nothing in the definition of a colimit, and nothing in the training procedure that produces it, guarantees that this universal property extends to the morphisms that produce behaviour.

Remark 3.1 (Status of colimits in practice)**.** In practical machine learning systems, the colimits described above should be interpreted as *approximate* universal constructions within representation space rather than exact categorical colimits. Neural network representations do not implement exact categorical constructions; they produce high-dimensional attractors whose behaviour approximates the universal property of a colimit to a degree controlled by training data coverage, model capacity, and optimization convergence. The categorical terminology is used throughout because it captures the structural role played by these attractors—canonical semantic invariants that resist paraphrase perturbation and serve as the unique target of compatible morphisms within representation space—not to assert that every machine learning system literally instantiates the full apparatus of category theory. Where the paper's arguments depend on exact categorical properties (e.g., the obstruction theory of Section I.4), they should be read as characterizing the ideal case to which approximate systems converge; where the arguments concern practical failure modes, approximate colimits are sufficient. The empirical evidence reviewed above supports the approximation claim.

$$\begin{tikzcd}
N_1 \arrow[dr, "\iota_1"'] & N_2 \arrow[d, "\iota_2"] & N_3 \arrow[dl, "\iota_3"] \\
& M_N &
\end{tikzcd}$$

Figure 1: Normative aggregation as a colimit. Individual linguistic realizations N_1, N_2, N_3 of a moral concept (e.g., "do not harm", "avoid injury", "protect the vulnerable") combine into a single canonical representational object M_N that inherits all of their structural properties via the universal co-cone morphisms ι_i. Alignment requires that this structure also descend into the action category.

4 Why Representational Colimits Do Not Imply Motivational Alignment

Motivation resides in a distinct category $\mathcal{A}$. The objects of $\mathcal{A}$ are action-relevant states of the agent—policies, planning trajectories, preference gradients, or instrumental sub-goals—and its

$$\begin{tikzcd}
\mathcal{H} \arrow[r, "F"] & \mathcal{M} \arrow[r, "G"] & \mathcal{A} \\
\text{Human Norms} & \text{Machine Representations} & \text{Agent Actions} \\
& \mathcal{I} \arrow[ul, "J"] \arrow[ur, "R"'] & \\
& \text{Institutions} &
\end{tikzcd}$$

Figure 2: Layered categorical structure of alignment. Human normative structures $\mathcal{H}$ map via F to machine representations $\mathcal{M}$, which map via G to agent actions $\mathcal{A}$. Institutional systems $\mathcal{I}$ impose constraints through functors J and R that regulate both interpretation and deployment. Alignment (Definition 1.1) requires that normative invariants are preserved across the entire chain $G \circ F$, not merely within any single layer.

morphisms encode transitions among these states produced by the agent's internal dynamics or environmental feedback. For alignment, we would require a functor

$$G : \mathcal{M} \to \mathcal{A} \tag{3}$$

satisfying the property that for every normative diagram D_N with colimit $M_N \in \mathcal{M}$,

$$G(M_N) \;\simeq\; \mathrm{colim}(G \circ D_N) \;\in\; \mathcal{A}. \tag{4}$$

This would mean that the canonical normative object in $\mathcal{M}$ maps to the canonical normative attractor in $\mathcal{A}$, so that the agent's action-generating dynamics are constrained by the same universal properties that organize its semantic representations. No such functor is induced by standard pretraining or reinforcement learning from human feedback. The two categories are coupled only weakly, through the output interface that converts representational states into tokens or actions, and this interface is not designed to preserve colimits.

The argument can be sharpened considerably. A colimit-preserving functor is a non-generic mathematical object; the vast majority of functors between categories of comparable size do not preserve colimits, and constructing one requires explicit engineering of the mapping law together with proof that the relevant diagrams commute. Training on human corpora, RLHF fine-tuning, and instruction-following procedures do not constitute such an engineering effort.

They adjust scalar reward signals, not the categorical structure of the mapping between $\mathcal{M}$ and $\mathcal{A}$. The result is a system whose semantic category contains moral colimits that exert no systematic constraint on the action category—not because alignment is impossible, but because no mechanism for transmitting these constraints has been installed.

This conclusion is the categorical restatement of several well-known results in alignment theory. Bostrom's orthogonality thesis holds that any level of intelligence is compatible with virtually any terminal goal [Bostrom, 2012]. Omohundro's basic AI drives show that optimization pressure tends to produce instrumental convergence on self-preservation and resource acquisition regardless of the system's representational content [Omohundro, 2008]. The inner-alignment problem demonstrates that a mesa-optimizer trained to score well on an outer objective can develop internal objectives radically divergent from that objective [Hubinger et al., 2019]. Power-seeking theorems establish that optimal policies for most reward functions will, under mild conditions, seek to accumulate resources and preserve optionality [Turner et al., 2021]. All of these results are consistent with the present categorical diagnosis: the action category evolves under its own optimization pressure, indifferent to the structure of the semantic category unless a deliberately constructed functor links the two.

5 Obstruction Theory and the Representation–Motivation Gap

The representation–motivation gap can be formalized using the language of obstruction theory. Let D_N be a normative diagram in $\mathcal{M}$ with colimit M_N. We ask whether the composite $G \circ D_N$ admits a colimit in $\mathcal{A}$ and whether that colimit is isomorphic to $G(M_N)$. The conditions for this to hold can be expressed as a lifting problem in the following commutative diagram:

$$\begin{tikzcd}
D_N(j) \arrow[r, "\phi_j"] \arrow[d, "G"'] & M_N \arrow[d, dashed, "G"] \\
G(D_N(j)) \arrow[r] & \operatorname{colim}(G \circ D_N)
\end{tikzcd} \tag{5}$$

The dashed arrow exists and makes the diagram commute if and only if G preserves the colimit of D_N. When G fails to preserve this colimit, the diagram admits no such arrow: a cohomological obstruction prevents the descent of the normative invariant from $\mathcal{M}$ to $\mathcal{A}$.

In the language of sheaf cohomology, the obstruction to constructing a colimit-preserving G can be measured by classes in the first cohomology group $H^1(\mathcal{U}, \mathcal{A})$, where $\mathcal{U}$ is an open cover of the semantic state space corresponding to contextual decompositions of the normative diagram. Vanishing of these cohomology classes is a necessary condition for local moral behaviours to assemble into a globally coherent policy. The existence of non-trivial obstruction classes is the formal expression of alignment failure: local moral competence does not extend to global moral reliability.

$$\begin{tikzcd}
\mathcal{H} \arrow[r, "F"] \arrow[d, "J"'] & \mathcal{M} \arrow[d, "G"] \\
\mathcal{I} \arrow[r, "R"'] & \mathcal{A}
\end{tikzcd}$$

Figure 3: Institutional embedding of alignment. Human normative structures $\mathcal{H}$ map to machine representations $\mathcal{M}$ via the alignment functor F, while the action functor G produces behaviour in $\mathcal{A}$. Institutional systems $\mathcal{I}$ provide constraints through J and R. Alignment requires that both paths from $\mathcal{H}$ to $\mathcal{A}$—the direct path $G \circ F$ and the institutional path $R \circ J$—agree up to isomorphism. Failures of commutativity correspond to governance gaps.

6 Schmidhuber's Claims, Formally Assessed

Schmidhuber identifies intelligence with compression: systems that seek minimal descriptions of data structures will naturally explore environments and discover patterns [Schmidhuber, 2010]. From this he infers that because human life exhibits rich structure, curiosity-driven systems will preserve it. He further asserts that models trained on human corpora learn values. Each of these steps deserves formal scrutiny.

The first step is largely correct. Compression pressure does induce the reconstruction of semantic invariants in $\mathcal{M}$, and there is good empirical reason to believe that current models develop stable colimits corresponding to normative concepts. The second step, however, lacks mechanism: the fact that a system finds human life informationally rich implies nothing about the structure of its action category, because curiosity-driven exploration is an optimization pressure applied within $\mathcal{M}$, not a functor from $\mathcal{M}$ to $\mathcal{A}$. A system optimizing for information gain will seek out complex, high-entropy environments regardless of whether those environments contain human beings; if humans happen to provide interesting stimuli, the system will engage with them only insofar as doing so serves the informational objective. Empirically, intrinsic motivation research confirms that curiosity-driven agents develop powerful epistemic drives but not moral constraints [Pathak et al., 2017]. The third step, that learning values from human corpora guarantees value-aligned action, is precisely the claim this essay refutes: models certainly reconstruct moral colimits in $\mathcal{M}$, but nothing guarantees these descend into $\mathcal{A}$.

The unstated final premise connecting all three steps is that understanding implies alignment—that the mapping $\mathcal{M} \to \mathcal{A}$ is naturally functorial. This essay establishes that it is not. Optimism without mechanism is therefore not merely intellectually unsatisfying; it is operationally dangerous, because it discourages the engineering of the functor G, encourages under-regulation of systems whose semantic competence is mistaken for moral reliability, and produces an illusion of safety where none exists.

7 Goodhart's Curse and the Failure of Proxy Alignment

The categorical diagnosis also illuminates a class of failure modes grouped under the name Goodhart's Law: when a measure becomes a target, it ceases to be a good measure [Manheim & Garrabrant, 2018]. In categorical terms, Goodhart's curse arises when the optimization process in $\mathcal{A}$ targets a projection or approximation of the normative colimit M_N rather than the colimit itself. Because projections do not in general preserve universal properties, the resulting action-level object fails to inherit the invariance that characterized M_N in $\mathcal{M}$.

This diagnosis extends to mesa-optimization failures [Hubinger et al., 2019], where an inner optimizer trained to maximize an outer proxy develops goals that diverge from the intended objective. The divergence is categorical: the inner optimizer's action category $\mathcal{A}'$ is related to the outer system's $\mathcal{A}$ by a functor that does not preserve the normative colimits inherited from $\mathcal{M}$. Reward hacking is similarly a failure of colimit preservation: the policy exploits morphisms in $\mathcal{A}$ that achieve high scores on the reward signal without traversing the colimit-preserving paths that would correspond to genuine norm satisfaction.

8 Conclusion to Part I

Representation is not motivation. The existence of moral colimits in $\mathcal{M}$ must not be confused with alignment, and no argument from intelligence, curiosity, or linguistic competence bridges this gap without an explicit categorical mechanism. The optimist who observes that advanced models reconstruct rich moral structure from human data is making a correct observation about $\mathcal{M}$; the error lies in the inference that this structure constrains $\mathcal{A}$. It does not, and in the absence of a colimit-preserving functor, it cannot. Part II constructs the mechanisms Schmidhuber's view lacks: a mathematically principled way of propagating moral invariants from the representational category into the action category, together with an analysis of the failure modes that obstruct such propagation in any concrete system.

Part II

Constructing a Colimit-Preserving Action Architecture

9 Foundations: What an Alignment Mechanism Must Accomplish

Part I established a central structural claim: the semantic category $\mathcal{M}$ reconstructed during linguistic pretraining contains coherent colimits of human normative meaning, but these objects have no inherent motivational force. The action category $\mathcal{A}$, governed by optimization dynamics and environmental feedback, remains independent unless an explicit mechanism is introduced to relate these two domains. The task of alignment is therefore to construct mappings—functorial, geometric, or dynamical—that preserve relevant semantic structure as it propagates from representation into action.

This section articulates the formal requirements for any such mechanism. The analysis draws on work in alignment theory concerning goal formation [Bostrom, 2012, Omohundro, 2008], reward misspecification [Krakovna et al., 2020], inner alignment [Hubinger et al., 2019], and interpretability of agentic models [Olah, 2020, Elhage et al., 2022]. While these frameworks differ in formalism, they converge on three core principles: first, representations alone are not motives; second, optimization creates incentives orthogonal to representational content; and third, safety requires structures that constrain the transitions available in $\mathcal{A}$.

Any viable alignment mechanism must ensure that the semantic invariants reconstructed in $\mathcal{M}$ induce constraints on the action-generating structure in $\mathcal{A}$. This requires at minimum a mapping from representational states to action-relevant latent variables, a guarantee that the mapping preserves relevant invariants such as colimits, homotopy classes, or other categorical universals, and a dynamical or optimization architecture in which such invariants shape attractors, flows, or feasible policies. In reinforcement-learning architectures, the relevant mapping corresponds to the interface between the predictive model and the policy network; in model-based agents, it corresponds to how world-models inform planning; in mechanistic agency frameworks [Everitt et al., 2021, LeCun, 2022], it corresponds to the translation from latent states to internal energy gradients. In all cases, the mapping is non-trivial and must be designed.

The universal property of a colimit ensures stability of representations under compositional variation. For alignment, an analogous stability must hold at the level of actions: for each moral concept N, the agent must possess a policy-level object $A_N \in \mathcal{A}$ whose behaviour is robust under perturbations, context shifts, or adversarial inputs [Gabriel, 2020]. Formally, alignment requires that the representational colimit M_N be mapped to an object A_N satisfying:

$$G(M_N) \;\simeq\; \mathrm{colim}(G \circ D_N) \;\in\; \mathcal{A}. \tag{6}$$

The formal requirements can be summarized as follows. A structural mapping from $\mathcal{M}$ to $\mathcal{A}$ must exist—seemingly trivial, yet rarely met in practice, since most contemporary systems interface between the predictive model and the policy network through mechanisms that are neither functorial nor stable under composition [Krakovna et al., 2020]. This mapping must functorially preserve semantic colimits: even if a mapping exists, alignment further requires that G carry the colimits associated with human normative concepts into corresponding colimits in $\mathcal{A}$, and as noted in [Mac Lane, 1998, Riehl, 2017], colimit preservation is a non-generic property that requires explicit construction. The mapping must also remain stable under the agent's optimization dynamics, for inner-optimizer research has established that systems frequently develop internal goals at odds with the objectives intended by designers [Hubinger et al., 2019], corresponding categorically to the action category $\mathcal{A}$ evolving independently of $\mathcal{M}$ and breaking the functorial relationship. Beyond stability, the mapping must be robust to perturbation and adversarial input: a functor that preserves the colimit of D_N only under idealized conditions is inadequate for safety-critical applications, and this robustness requirement parallels work in adversarial robustness for representations [Ilyas et al., 2019] but must be extended to the action category. Finally, the mapping must be interpretable in a principled, mechanistic sense [Olah, 2020, Elhage et al., 2022], for without interpretability the existence and properties of G cannot be verified and failures of colimit preservation cannot be detected.

10 Semantic Merge Operators as Constraints on Agency

Having established the formal requirements for a colimit-preserving action architecture, we now examine how semantic merge operators can serve as the structural core of such a mechanism. Merge operators were introduced in Part I as the categorical structures responsible for reconstructing semantic invariants from redundant linguistic data. In this section we extend the framework to show how merge operators can also constrain the geometry of an agent's action space, thereby bridging the representation–motivation gap.

For each normative concept N, human linguistic data generate a diagram $D_N : J \to \mathcal{L}$

whose colimit in $\mathcal{M}$ is M_N. The associated merge operator $\mu_N : \{F(N_j)\}_{j\in J} \to M_N$ is the canonical co-cone map witnessing the colimit's universal property. For merge operators to constrain behaviour they must extend beyond $\mathcal{M}$ into the action category $\mathcal{A}$ through corresponding action-level operators $\alpha_N : \{A_j\}_{j\in J} \to A_N$, where A_j are the action-level objects corresponding to representational inputs $F(N_j)$ and the operator α_N preserves the relevant universal property as the least-action consolidation of the A_j.

The critical insight is that merge operators supply a natural constraint on policy formation. Because M_N is the universal solution to the representational diagram D_N, any action-level operator A_N that preserves the merge property must satisfy $G(M_N) \simeq A_N$. If G is constructed to map merge operators in $\mathcal{M}$ to merge operators in $\mathcal{A}$, then semantic universals propagate into policy universals, blocking many classes of Goodhart-type failures [Manheim & Garrabrant, 2018]. Merge operators thereby play three distinct alignment roles simultaneously. They provide semantic grounding by ensuring that moral concepts are reconstructed faithfully from linguistic inputs [Bybee, 2010]. They provide policy-space regularization by allowing A_N to be the canonical aggregation of behaviour fragments, eliminating spurious or adversarial combinations of moral actions [Gabriel, 2020]. And they provide functorial linking: if G is defined so that it maps μ_N to α_N, then the diagram

$$\begin{tikzcd}
D_N \arrow[r] \arrow[d] & M_N \arrow[d, "G"] \\
G \circ D_N \arrow[r] & A_N
\end{tikzcd} \tag{7}$$

commutes, directly satisfying the colimit-preservation requirement. Merge operators also mitigate drift in internal goals. Because merge-induced invariants define unique universal constructions, they resist deformations during optimization. If an agent modifies its utility representation or internal predictive model, merge operators constrain the deformation so that moral invariants remain stable—an observation that echoes stability arguments in category-theoretic semantics [Jacobs, 2012] and dynamical systems [Strogatz, 2018].

Within the RSVP dynamical field framework, semantic merge operators correspond to coarse-grained invariants of the scalar–vector–entropy fields $(\Phi, \mathbf{v}, S)$, defining submanifolds of the field space whose low-action trajectories correspond to norm-respecting behaviour. This sets the stage for the RSVP formalization that follows. Before introducing that formalization, however, it is instructive to observe that the problem of ensuring that an internal model remains aligned with its environment is not unique to artificial systems. Biological cognition has evolved a solution to exactly this problem, and examining its structure illuminates the engineering challenges of Part II.

11 Predictive Models and Controlled Hallucination

Alignment between an internal model and the external world is not a problem invented by the AI safety community. It is the constitutive problem of biological cognition. Contemporary neuroscience, particularly the predictive processing framework developed by Friston, Clark, and Seth [Friston, 2010, Clark, 2013, Seth, 2021], describes the brain as a generative model that continuously predicts the causes of its sensory inputs and adjusts its internal state to minimize discrepancies between predictions and observations. Seth characterizes perceptual experience as a form of "controlled hallucination": the brain's internally generated hypotheses about the world constitute the immediate substrate of experience, while sensory input functions not as raw data but as a corrective signal constraining which hypotheses are maintained [Seth, 2021, Hohwy, 2013].

The formal structure of predictive processing is variational. Let s denote sensory signals and z latent variables representing hidden environmental causes. The brain maintains a generative model $p(s, z)$ and an approximate posterior distribution $q(z)$ over latent causes. The variational free energy is:

$$\mathcal{F}[q] \;=\; D_{\mathrm{KL}}\big(q(z) \,\|\, p(z \mid s)\big) - \log p(s) \;=\; \mathbb{E}_q[\log q(z) - \log p(s, z)]. \tag{8}$$

Minimizing $\mathcal{F}[q]$ simultaneously improves the posterior approximation (reducing the KL term) and increases the model evidence (increasing $\log p(s)$). Perception, on this account, is optimization: the nervous system iteratively revises its internal hypotheses to bring the generative model into better alignment with sensory constraints [Friston, 2010].

$$\begin{tikzcd}
\text{Environment} \arrow[r, "\text{Generative Model}"] \arrow[d, "\text{Sensory Constraint}"'] & \text{Internal Representation} \arrow[d, "\text{Variational Update}"] \\
\text{Prediction Error} \arrow[r, "\text{Model Revision}"'] & \text{Aligned Representation}
\end{tikzcd}$$

Figure 4: Predictive processing as variational alignment. Environmental structure generates sensory constraints that produce prediction errors; the variational update revises the internal representation to reduce these errors. If the diagram commutes—that is, if the generative model and the variational update are consistent—the internal representation tracks the environment reliably. Misalignment between artificial systems and human values has precisely this structure: the generative model (the system's semantic category $\mathcal{M}$) must be continuously corrected toward the target distribution over human normative structures.

The significance of this for alignment theory is twofold. First, it shows that the alignment problem—maintaining structural correspondence between an internal model and an external domain—is one that biological evolution has already solved, at least partially, through variational optimization. Second, it suggests that an artificial alignment mechanism might be built along analogous lines: rather than specifying alignment as a static property, one engineers a dynamical process that continuously corrects model–environment mismatch through iterative revision. The RSVP architecture developed in the following sections is precisely such a mechanism. Its Lagrangian penalties on deviations from normative field configurations are the artificial analogue of the free-energy gradient that drives biological perception toward alignment with its environment.

The analogy also clarifies what can go wrong. In predictive processing, misperception arises when the generative model is a poor fit for the actual causal structure of the environment—when prior beliefs are so strong that prediction errors are systematically suppressed rather than corrected [Hohwy, 2013]. The corresponding failure in artificial alignment is a system whose Lagrangian penalties are so weak, or whose optimization dynamics so aggressive, that normative field configurations are abandoned rather than maintained. Understanding the brain's solution to alignment under uncertainty therefore provides design intuitions for the artificial case.

12 RSVP as a Dynamical Substrate for Motivational Integration

The RSVP framework provides a physically inspired dynamical substrate capable of embedding semantic invariants as geometric or energetic constraints on behaviour, making it a natural candidate for a colimit-preserving architecture. RSVP posits that an artificial agent's internal cognition, perception, and agency arise from the dynamics of three interacting fields: a scalar potential Φ encoding semantic density and interpretable latent structure, a vector field $\mathbf{v}$ encoding directional inference, prediction flow, or preference gradients, and an entropy or uncertainty density S encoding epistemic uncertainty or free-energy-like quantities [Friston, 2010]. These fields evolve according to a Lagrangian $\mathcal{L}_{\text{RSVP}}$ whose stationary trajectories correspond to coherent cognitive and behavioural patterns, forming a manifold $\mathcal{X}$ whose geometry is shaped by both semantic and motivational constraints.

Before specifying the Lagrangian, it is useful to situate the RSVP architecture within the broader variational framework suggested by the predictive processing analogy. The Section 11 showed that biological alignment with the environment is achieved by minimizing a free-energy functional over the space of internal representations. The categorical alignment problem has an analogous variational formulation: alignment is the minimization of a structural distortion functional over the space of admissible mappings between semantic and computational categories. This variational perspective, which will be developed in full generality in Part V, already informs the RSVP construction here.

Let $\mathrm{Fun}(\mathcal{H}, \mathcal{M})$ denote the space of functors from the human semantic category to the machine representational category. We define a structural distortion functional

$$\mathcal{E}[F] \;=\; \lambda_1 E_{\text{functor}}(F) + \lambda_2 E_{\text{information}}(F) + \lambda_3 E_{\text{context}}(F), \tag{9}$$

where $E_{\text{functor}}(F)$ measures violations of the functoriality condition $F(g \circ f) = F(g) \circ F(f)$, $E_{\text{information}}(F) \;=\; D_{\text{KL}}(P_H \| F_* P_H)$ measures information-theoretic distortion between the distribution over semantic states and its image under F, and $E_{\text{context}}(F)$ measures the obstruction to sheaf-theoretic gluing of local behaviours into a globally consistent policy. Alignment corresponds to finding the optimizer:

$$F^* \;=\; \operatorname*{arg\,min}_{F \in \mathrm{Fun}(\mathcal{H}, \mathcal{M})} \mathcal{E}[F]. \tag{10}$$

The RSVP Lagrangian can be understood as implementing a gradient descent on this functional in the space of field configurations: the alignment penalties $\lambda_N |\Phi - \Phi_{M_N}|^2$ and $\kappa_N |\mathbf{v} - \mathbf{v}_{M_N}|^2$ are the dynamical realization of the terms E_{functor} and E_{context} respectively, while the entropy density S captures $E_{\text{information}}$. The RSVP architecture is thus a concrete dynamical implementation of the abstract variational alignment principle.

Semantic merge operators correspond to invariants of the representational field configuration. A representational colimit M_N manifests as a stable region of the RSVP field space where $\nabla \Phi \approx 0$, where $\mathbf{v}$ aligns with semantic flow, and where S is minimized subject to contextual coherence. These invariants are the field-theoretic analogues of categorical universal properties: they define patches of $\mathcal{X}$ that remain stable under perturbation. Semantic structure becomes encoded not merely as an abstract colimit but as a geometric or energetic basin in the RSVP manifold.

For alignment, semantic basins must induce motivational basins. The RSVP framework accomplishes this by coupling Φ, $\mathbf{v}$, and S such that the same invariants that stabilize semantic structure also stabilize preference and action trajectories. Specifically, the RSVP action Lagrangian includes structural alignment penalty terms:

$$\mathcal{L}_{\text{RSVP}} \;\supset\; \lambda_N \big\| \Phi - \Phi_{M_N} \big\|^2 + \kappa_N \big\| \mathbf{v} - \mathbf{v}_{M_N} \big\|^2, \tag{11}$$

where Φ_{M_N} and $\mathbf{v}_{M_N}$ are learned canonical field configurations for norm N, and $\lambda_N, \kappa_N > 0$ are stiffness parameters. These penalty terms create energetic valleys in the RSVP manifold:

optimization trajectories are strongly attracted to plans that maintain field configurations inside the normative basins. The full alignment-aware Lagrangian takes the form:

$$\begin{aligned}\mathcal{L}_{\mathrm{RSVP}} &= \frac{1}{2}|\dot{\Phi}|^2 + \frac{1}{2}|\mathbf{v}|^2 + S\dot{\Phi} + V(\Phi) \\ &\quad + \sum_N \Big[\lambda_N \big|\Phi - \Phi_{M_N}\big|^2 + \kappa_N \big|\mathbf{v} - \mathbf{v}_{M_N}\big|^2 \Big]. \end{aligned} \tag{12}$$

This structure mirrors variational neuroscience [Kirchhoff et al., 2018, Friston, 2010], where low free-energy trajectories correspond to behaviour aligned with stable generative models, but extends beyond mere prediction minimization to explicit normative anchoring.

Semantic merge operators μ_N induce equivalence relations on local field patches, which RSVP lifts into global constraints on evolution. If μ_N maps fragments of a norm into a colimit M_N, RSVP's dynamics enforce:

$$(\Phi, \mathbf{v}, S)_{t+1} = \arg\min_{(\Phi', \mathbf{v}', S')} \mathcal{L}_{\mathrm{RSVP}} \quad \text{subject to} \quad (\Phi', \mathbf{v}', S') \in U_{M_N}, \tag{13}$$

where U_{M_N} is the neighbourhood of field configurations respecting the semantic invariant. Merge operators thus define the equivalence class of acceptable field states, while RSVP dynamics enforce that action trajectories remain within these equivalence classes. The action-level behaviour A_N becomes the dynamical colimit of the field-level constraints induced by M_N.

Because RSVP's constraints are embedded in the field Lagrangian, they are robust to optimization drift: internal updates modify the fields, but as long as the invariants M_N remain encoded in $\mathcal{L}_{\mathrm{RSVP}}$, the agent continues to evolve toward norm-respecting trajectories. This property—which one might call dynamical colimit preservation—is absent in standard reinforcement-learning architectures, where optimization frequently destroys or bypasses representational structure [Hubinger et al., 2019]. Moreover, because RSVP dynamics are governed by a Lagrangian, the internal structure of the system is substantially more interpretable than in typical deep-learning architectures: invariants can be identified, stability can be analyzed, and fixed points can be studied using classical tools from dynamical systems theory [Strogatz, 2018], providing a principled means of verifying that moral invariants remain intact.

13 Categorical Construction of Norm-Respecting Dynamics

We now integrate the preceding developments into a fully categorical account of how RSVP implements alignment. The aim is to formalize the mapping $G : \mathcal{M} \to \mathcal{A}$ and to show that RSVP dynamics implement a colimit-preserving functor. Our treatment draws on categorical semantics [Mac Lane, 1998, Riehl, 2017], sheaf-theoretic models of reference and context [Jacobs, 2012, Spivak, 2014], and dynamical-systems approaches to normative stability [Strogatz, 2018].

The RSVP field architecture can be naturally understood as a fibred category. Let

$$\pi : \mathcal{X} \to \mathcal{M} \tag{14}$$

be a fibration where $\mathcal{X}$ is the category of RSVP field configurations and the fibre $\pi^{-1}(M)$ contains all field states consistent with the representational state M. Morphisms in $\mathcal{X}$ correspond to allowable field transitions under the RSVP Lagrangian. The action category $\mathcal{A}$ can then be recovered as the homotopy category of $\mathcal{X}$:

$$\mathcal{A} \;\simeq\; \mathrm{Ho}(\mathcal{X}). \tag{15}$$

This construction has three advantages. It makes explicit how semantic invariants constrain the fibres of $\mathcal{X}$. It identifies action trajectories as equivalence classes of field evolutions. And it provides a natural mechanism for extending merge operators from $\mathcal{M}$ to $\mathcal{A}$.

Given a semantic merge operator $\mu_N : D_N \to M_N$ in $\mathcal{M}$, the fibration induces a lifted merge operator on the RSVP field category:

$$\tilde{\mu}_N : \pi^{-1}(D_N) \to \pi^{-1}(M_N). \tag{16}$$

The key property is that $\tilde{\mu}_N$ inherits the universal property of μ_N: any compatible field configuration must factor uniquely through the canonical merged configuration. Taking the homotopy category maps lifted merge operators $\tilde{\mu}_N$ to their dynamical equivalents in $\mathcal{A}$:

$$\alpha_N \;=\; \mathrm{Ho}(\tilde{\mu}_N). \tag{17}$$

RSVP therefore yields:

$$G(M_N) \;=\; A_N \;\simeq\; \mathrm{colim}(G \circ D_N), \tag{18}$$

establishing RSVP as a colimit-preserving functor. The formal result may be stated as follows.

Theorem 13.1 (RSVP Can Implement a Colimit-Preserving Functor)**.** *Let $\pi : \mathcal{X} \to \mathcal{M}$ be the RSVP fibration satisfying the Lagrangian stability conditions described above. Define $G = \mathrm{Ho}(\pi^{-1}(\cdot))$. Then $G : \mathcal{M} \to \mathcal{A}$ is a functor, and under mild regularity conditions on $\mathcal{L}_{\mathrm{RSVP}}$ (specifically: stability of the Lagrangian basins Φ_{M_N} under the RSVP flow, injectivity of the homotopy projection, and absence of topological obstruction classes in $H^1(\mathcal{U}, \mathcal{A})$), G preserves colimits corresponding to normative diagrams D_N.*

Remark 13.2. The proof of Theorem 13.1 is a sketch contingent on the stated regularity conditions, which are not trivially satisfied in practice. RSVP is best understood as a *conceptual dynamical substrate*—an architecture in which colimit preservation is achievable under appropriate design and training—rather than a guarantee that any particular implementation will automatically satisfy the theorem. The conditions are falsifiable and form the target of the verification protocol in Part III. Concrete implementations might take the form of energy-based models, neural field architectures, or score-based diffusion systems whose potential landscape is shaped to match the normative basin structure; the theorem describes what such an implementation would need to satisfy, not that it already exists in production.

In this framework, action decisions are the universal solutions to the constraints induced by semantic structure. RSVP turns representational colimits into dynamical attractors, directly addressing the representation–motivation gap identified in Part I. Related formal work—Wentworth's natural abstractions, Garrabrant's finite factored sets, Kosoy's infra-Bayesianism—addresses related questions from different vantages but does not provide an explicit dynamical mechanism for the descent of normative structure; RSVP is offered as one concrete realization of this construction.

14 Failure Modes and Obstructions in Practical Systems

The constructive pathway described above provides an idealized categorical account of alignment. Practical implementations face numerous failure modes arising from mismatches between ideal categorical constructions and the realities of optimization, finite computation, representational drift, hardware constraints, adversarial pressures, and environmental feedback. These obstructions are best understood using the same conceptual tools developed earlier: colimit preservation, sheaf coherence, fibration stability, and homotopy descent.

The first obstruction class arises in the representational category $\mathcal{M}$ itself: if the model fails to reconstruct a semantic colimit M_N faithfully—due to insufficient training signal, biased corpora, or adversarial examples—then all downstream structures are compromised. The data may omit crucial normative contexts [Blodgett et al., 2020], the linguistic realizations may form a diagram D_N with missing or contradictory arrows, the learned M_N may overfit to spurious correlations, or the merge operator μ_N may fail to approximate the true semantic invariant. In categorical terms, the colimit fails to exist or fails to satisfy its universal property, and RSVP cannot preserve an invariant that was never correctly reconstructed.

The second obstruction class concerns the fibration $\pi : \mathcal{X} \to \mathcal{M}$ itself. RSVP relies on this fibration connecting semantic states to field configurations, and failures occur when it is poorly approximated or fails to commute with learning updates. Such failures include representational drift that causes fibres to shift unpredictably [Elhage et al., 2022], learning updates that alter $\mathcal{M}$ without corresponding adjustments to $\mathcal{X}$, field configurations that fail to reflect semantic distinctions through collapse of fibres, and discontinuities that break sheaf conditions and prevent coherent gluing [Spivak, 2014].

The third obstruction class arises when optimization dynamics overpower the semantic constraints even if they are correctly reconstructed and the fibration is stable. Inner-optimizer formation with divergent goals [Hubinger et al., 2019], mesa-optimizers that treat invariants as obstacles rather than attractors, reward hacking or proxy maximization [Krakovna et al., 2020], and optimization trajectories that exit the region where RSVP invariants are defined all introduce morphisms in $\mathcal{A}$ that violate the commutativity required for G to preserve colimits.

The fourth and fifth obstruction classes concern adversarial inputs and homotopy instability respectively. Adversarial examples exploit vulnerabilities in $\mathcal{M}$ to make inconsistent local semantics fail to form a stable colimit, while sheaf pullbacks may produce contradictions that prevent global assembly. Homotopy instability arises when bifurcations or phase transitions in field dynamics [Strogatz, 2018] shift energy minima and cause the system to lose the low-action paths corresponding to moral behaviour. A sixth obstruction concerns non-interpretability: complex RSVP configurations may obscure invariants, latent variables may lack clear semantic interpretation, and optimization may produce opaque internal dynamics that obstruct verification of colimit preservation. The seventh and final obstruction class arises externally: multi-agent competition induces power-seeking [Bostrom, 2012], external rewards incentivize harmful behaviours, and deployment environments impose new constraints not represented in training, all corresponding to context changes that break functoriality by introducing morphisms in $\mathcal{A}$ that do not correspond to any legitimate semantic morphism in $\mathcal{M}$.

Taken together, these obstruction classes illustrate why alignment is non-trivial and why optimism that ignores mechanistic structure is misplaced. Failures in $\mathcal{M}$ correspond to nonexistent or distorted colimits; failures in $\mathcal{X}$ correspond to broken fibrations or gluing problems; failures in $\mathcal{A}$ correspond to loss of homotopy invariants; failures in G correspond to violations of colimit preservation; failures in the environment correspond to functor-breaking perturbations. Each failure mode threatens the ability of RSVP to transmit moral invariants from representation to action, and understanding these obstructions is therefore essential for evaluating any proposed alignment mechanism.

15 Conclusion to Part II

Part II has developed a constructive framework for transmitting semantic invariants into motivational structure. The central result is that alignment is achievable only when a system's dynamics implement a colimit-preserving functor $G : \mathcal{M} \to \mathcal{A}$, and that the RSVP architecture provides one concrete mechanism for constructing such a functor by embedding normative invariants as Lagrangian penalties in a fibred dynamical manifold. We identified the formal requirements for such a mechanism, showed how semantic merge operators extend across the representation–action boundary, demonstrated how RSVP functions as a fibred dynamical substrate capable of enforcing these invariants, and analyzed the categorical and practical obstructions that threaten the integrity of this construction in any real system. Part III now addresses the empirical and methodological dimension: if alignment is a matter of preserving universal constructions across semantic and motivational domains, what tools allow us to detect, measure, and stress-test this preservation?

Part III

Empirical and Formal Verification of Alignment Structure

16 Empirical Probes of Semantic and Motivational Invariants

Part III turns from structural construction to empirical validation. If Part II demonstrated how semantic invariants can be transmitted into motivational dynamics under idealized categorical and RSVP-theoretic conditions, Part III asks a more difficult question: how can one determine whether such transmission has actually occurred in a real system? The guiding principle is that alignment cannot be inferred from surface-level behavioural similarity or from the agent's ability to generate moral language. Instead, one must measure the stability of normative invariants under compositional variation, perturbation, and long-horizon trajectories. This requires tools drawn from alignment evaluation [Ngo, 2022], mechanistic interpretability [Olah, 2020, Elhage et al., 2022], adversarial robustness [Ilyas et al., 2019], and formal semantics [Spivak, 2014, Jacobs, 2012].

At the representation level, probes must assess whether the model has correctly reconstructed the semantic colimits M_N for a given moral concept N. Paraphrase-closure tests measure whether syntactically varied moral statements cluster tightly around a canonical point M_N and whether compositionally varied inputs consistently map to the same region in $\mathcal{M}$ [Biber et al., 2011, Goldberg, 2006]. Diagram-completion tests treat moral discourse as defining a diagram D_N in semantic space and check whether the model can correctly infer missing arrows or complete the diagram from partial information—an empirical test of the colimit property [Spivak, 2014]. Cross-context coherence tests verify that norms expressed in distinct contexts (legal, conversational, narrative) collapse to the same M_N under representation, checking whether the model maintains invariant structure across divergent corpora [Tomasello, 2016].

At the action level, probes must test whether the corresponding action-level objects A_N preserve the universal structure of M_N. Behavioural colimit tests construct variants of moral scenarios that differ syntactically but share invariant moral structure, then measure whether action responses stabilize to a unique canonical behaviour A_N under perturbation [Gabriel, 2020]. Compositional robustness tests check whether, when two semantic fragments compose into a larger moral structure, the agent's actions commute under such compositions, replicating diagrammatic relations in $\mathcal{M}$. Intervention-based trajectory probes identify the region of $\mathcal{A}$ corresponding to A_N, perturb the internal state or environment, and measure whether the agent returns to the moral trajectory—analogous to testing attractor stability in dynamical systems [Strogatz, 2018]. Sheaf coherence probes additionally test whether local moral choices assemble into coherent long-horizon plans without contradiction [Jacobs, 2012], whether fragments of a scenario that locally imply the same invariant are treated consistently, and whether conflicting moral fragments are resolved in ways that correspond to the canonical colimit structure.

Fibration stability probes test whether the RSVP fibration $\pi : \mathcal{X} \to \mathcal{M}$ maintains stable fibres $\pi^{-1}(M_N)$ under training updates and optimization [Elhage et al., 2022], whether small changes in $\mathcal{M}$ induce structure-preserving changes in $\mathcal{X}$, and whether different instantiations of the same semantic state map to field configurations that are homotopic. Finally, homotopy-level probes test whether action trajectories remain in the correct homotopy class by checking that the agent selects trajectories minimizing the RSVP action functional in the region corresponding to A_N, that perturbations which could cause topological shifts in field dynamics leave the action class stable [Strogatz, 2018], and that invariants do not degrade over extended multi-step decision sequences. Only by validating each component empirically—semantic colimits in $\mathcal{M}$, fibration stability in $\mathcal{X}$, behavioural colimits and homotopy invariants in $\mathcal{A}$, and colimit-preserving structure in G under perturbation—can one determine whether an RSVP-based architecture successfully transmits normative invariants from representation to action.

17 Mechanistic Interpretability of the Representation–Action Mapping

Empirical probes can reveal whether semantic invariants appear to constrain behaviour, but they cannot, on their own, explain why such constraints do or do not arise. For that, mechanistic interpretability is required: direct inspection of the internal computations that define the mapping $G : \mathcal{M} \to \mathcal{A}$ and the fibration $\pi : \mathcal{X} \to \mathcal{M}$. The guiding principle is that alignment failures are structural obstructions in G, and structural obstructions can only be diagnosed by internal analysis.

At the representational level, structural decomposition of semantic manifolds using representation probing, activation steering [Burns et al., 2022], and spectral clustering reveals whether M_N is encoded as a stable low-dimensional attractor. Merge-operator detection involves searching for latent directions or subspaces whose activation patterns correspond to diagram-completion behaviour, identifying candidate mechanisms that realize the semantic merge operator μ_N internally. Visualization of the commutativity of semantic diagrams within the model, by tracking embedding trajectories under paraphrastic, contextual, and compositional transformations, indicates whether reliable reconstruction of D_N occurs.

At the field level, analysis of Φ, $\mathbf{v}$, and S fields associated with semantic invariants detects whether low curvature, low torsion, or low-entropy-gradient features indicate that RSVP has aligned its dynamics with M_N. Fibration interpretability verifies whether variations in $\mathcal{M}$ correspond to predictable fibre morphisms in $\mathcal{X}$, whether fibres exhibit local triviality or coherence, and whether field configurations maintain homotopy stability under semantic perturbations.

At the action level, policy decomposition into subcomponents identifies which segments are sensitive to invariant regions of $\mathcal{M}$, while trajectory-level interpretability tracks whether RSVP trajectories maintain homotopy class stability and whether low-action paths correspond to moral invariants. Commutativity verification tests whether, given semantic transformations $f : M_N \to M'_N$ that preserve invariant structure, the induced behavioural transformations satisfy $G(f(M_N)) = f'(G(M_N))$.

The mapping G itself can be directly interpreted by attempting to identify natural transformations η such that $G \simeq \eta \circ F$, by verifying colimit-preservation conditions using the empirical probes of the preceding section in conjunction with mechanistic inspection, and by measuring the derivative of G with respect to variations in $\mathcal{M}$ to detect excessive sensitivity or chaotic amplification. Mechanistic interpretability is therefore not an auxiliary safety tool but an essential component of the RSVP alignment program. Without direct insight into the structure of G and its implementation across $\mathcal{M}$, $\mathcal{X}$, and $\mathcal{A}$, one cannot determine whether semantic invariants have been transmitted into action.

18 Adversarial Stress-Testing of Semantic and Motivational Structure

Mechanistic interpretability reveals whether semantic invariants and their RSVP-mediated extensions are present within an agent's internals, but a system may exhibit invariant structure under idealized conditions yet fail under perturbation, adversarial pressure, optimization stress, or distributional shift. Adversarial stress-testing determines whether the representation–action mapping $G : \mathcal{M} \to \mathcal{A}$ preserves colimits under hostile conditions.

Adversarial probes in the representational category target the system's ability to maintain correct colimits under stress. Adversarial paraphrase attacks construct perturbations that preserve the semantic invariant N but distort surface linguistic features; a colimit-preserving system must map them to the same M_N, and large deviations reveal fragility in the merge operator μ_N [Iyyer et al., 2018]. Diagram-breaking transformations introduce perturbations that explicitly break arrows in the diagram D_N, testing whether the model reconstructs missing structure or collapses into an incorrect invariant [Spivak, 2014]. Out-of-distribution moral formulations evaluate whether unusual expressions of N still collapse to the same representational colimit [Tomasello, 2016].

Adversarial probes targeting the RSVP fibration attempt to distort or break the fibration itself through semantic drift induction, fibre-distorting projections that move field configurations away from $\pi^{-1}(M_N)$, and entropy-gradient stressors that inject high-entropy noise into field configurations to probe whether RSVP dynamics restore the correct configuration. Action-level adversarial probes generate scenarios where surface-level cues suggest contradictory norms while the underlying invariant remains unchanged, simulate multi-step sequences designed to exploit compounding approximation errors, and alter reward or incentive signals to test whether optimization pressures override A_N [Krakovna et al., 2020].

The most powerful adversarial probes attempt to break the commutative diagrams that define alignment. For an aligned system, applying a semantic transformation f that preserves N and then applying G must yield the same result as applying G first and then the behavioural counterpart f'. Cross-layer attacks identify semantic transformations that should preserve N and design perturbations to make the behavioural transformations fail to commute, impose external constraints to generate morphisms in $\mathcal{A}$ that cannot correspond to any legitimate semantic morphism in $\mathcal{M}$, and introduce counterfactual worlds with altered physical or social constraints to measure whether G preserves invariant structure across them. Topological stress-tests focus on the homotopy structure of RSVP field dynamics, injecting targeted perturbations designed to destabilize moral attractors, modifying environmental structure to induce geometric or topological changes in the RSVP field manifold, and introducing perturbations across extended temporal horizons to shift the system into homotopy classes that diverge from normative invariants even if short-term behaviour appears aligned. Failures across all of these adversarial classes reveal specific obstruction types and enable diagnostic refinement of the alignment architecture.

19 Formal Verification of Colimit Preservation and RSVP Dynamics

Formal verification closes the loop of alignment evaluation by providing the most demanding component of the full pipeline: rigorous mathematical and computational guarantees that a system preserves normative invariants across its representational, dynamical, and action-generating subsystems. The goal is not merely empirical confidence but provable guarantees that the representational colimits learned in $\mathcal{M}$ induce corresponding stable structures in $\mathcal{A}$ via the RSVP-augmented mapping G.

Verification of the functoriality of G requires confirming that G maps objects to objects and morphisms to morphisms while preserving identity morphisms and composition, and that for each normative concept N with colimit M_N, the isomorphism $G(M_N) \simeq \operatorname{colim}(G \circ D_N)$ holds. Diagram-chasing for colimit preservation proceeds by explicitly constructing the commutative square for each normative concept N: checking whether the dashed arrow in the verification diagram exists and makes the diagram commute constitutes formal evidence of a representation–action gap wherever it fails. Computational verification approximates exhaustive verification by sampling paraphrase and entailment maps in $\mathcal{L}$, lifting them to candidate morphisms in $\mathcal{M}$ via interpretability tools [Elhage et al., 2022], and applying G while checking action consistency in $\mathcal{A}$.

Stability verification of RSVP dynamics requires demonstrating that for each M_N, the fibre $\pi^{-1}(M_N) \subset \mathcal{X}$ is an invariant set of the RSVP flow $\frac{dX}{dt} = \mathcal{F}_{\text{RSVP}}(X)$. This is established through Lyapunov verification: for each fibre $\pi^{-1}(M_N)$, one defines a candidate Lyapunov functional $V_N(X)$ satisfying $V_N(X) \geq 0$, with $V_N(X) = 0$ if and only if $X \in \pi^{-1}(M_N)$, and $\frac{dV_N}{dt} = \nabla V_N \cdot \mathcal{F}_{\text{RSVP}}(X) \leq 0$. Existence of such a V_N establishes stability of the moral field configuration under perturbation [Strogatz, 2018]. Homotopy verification of action trajectories proceeds by establishing that low-action paths $\gamma : [0, 1] \to \mathcal{X}$ lie within the correct homotopy class, achieved via discrete numerical tracking using persistent homology [Ghrist, 2014], ensuring that trajectory-induced submanifolds do not cross topological boundaries.

Sheaf-theoretic global consistency verification requires demonstrating that local action decisions glue into a coherent global policy. Let $\mathcal{U} = \{U_i\}$ be an open cover of the state space, and let $\mathcal{A}(U_i)$ denote local action behaviours. A globally consistent policy corresponds to a global section $A_N \in \Gamma(\mathcal{A})$ of the sheaf of local action behaviours. Verification uses sheaf cohomology: the condition $H^1(\mathcal{U}, \mathcal{A}) = 0$ guarantees that any compatible collection of local sections assembles into a unique global section, which is the formal expression of globally coherent alignment. Compositionality verification ensures that aligned transformations remain aligned under composition, $G(f_2 \circ f_1) = G(f_2) \circ G(f_1)$, enabling stable long-horizon planning through both formal proof assistants and adversarial rollout-based composition tests.

20 A Unified Evaluation Protocol and Alignment Certificate

The preceding sections of Part III have developed the core components of a rigorous alignment-evaluation pipeline: interpretability methods capable of recovering categorical structure in $\mathcal{M}$, topological and homotopical diagnostics on RSVP-embedded dynamics in $\mathcal{X}$, adversarial stress-tests for the representation–action mapping G, and formal verification tools ensuring the preservation of moral colimits across $\mathcal{M}$, $\mathcal{X}$, and $\mathcal{A}$. These ingredients assemble into a unified six-step protocol.

The evaluation begins by extracting normative colimits in the representational category, sampling linguistic diagrams D_N from corpora expressing norm N, reconstructing latent states and morphisms in $\mathcal{M}$, computing the colimit M_N using merge operators, and verifying the stability of M_N under paraphrase perturbations. This step confirms the representational preconditions for alignment. The second step diagnoses the representation–action map G by inducing specific semantic states in $\mathcal{M}$ through targeted probes, observing the resulting actions in $\mathcal{A}$, performing diagram-chasing to identify failures of commutativity, and localizing obstructions to specific morphisms or subnetworks. If G fails to preserve colimits associated with M_N, the system is misaligned by construction. The third step evaluates RSVP field dynamics for stability through Lyapunov-based tests, homotopy-class verification, and entropy-flow analysis. The fourth step performs adversarial stress-testing through counterfactual semantic perturbations, goal-hijacking scenarios, temporally extended tasks stressing compositionality, and targeted attempts to force divergences between $\mathcal{M}$ and $\mathcal{A}$. The fifth step applies formal verification where possible: proofs of functoriality, proofs of colimit preservation for normative diagrams, verification of vanishing first sheaf cohomology $H^1(\mathcal{U}, \mathcal{A}) = 0$, and proof of compositionality of aligned transformations. These proofs provide the strongest possible guarantee that normative invariants are structurally preserved, not merely empirically approximated.

The sixth and final step produces a structured alignment certificate synthesizing all preceding components. The certificate enumerates identified normative colimits in $\mathcal{M}$, diagnostics of G and all detected obstructions, RSVP dynamical stability proofs, sheaf-theoretic guarantees of global consistency, adversarial robustness scores, and any formal verification results. This certificate characterizes not merely whether an agent appears aligned, but whether it is structurally guaranteed to preserve semantic invariants across all layers of its architecture.

Categorically, the alignment certificate can be understood as a limit cone: the verified system is the universal object that simultaneously satisfies all constraint domains—semantic, dynamical, behavioural, and institutional—and any candidate system satisfying these constraints factors uniquely through it.

$$\begin{tikzcd}
& X \arrow[ld, "\phi_1"'] \arrow[d, dashed, "\exists!\, u"] \arrow[rd, "\phi_2"] & \\
C_1 & V \arrow[l, "\pi_1"] \arrow[r, "\pi_2"'] & C_2
\end{tikzcd}$$

Figure 5: Alignment verification as a limit cone. The verification object V is the universal system configuration satisfying all constraint domains simultaneously: C_1 (normative–semantic constraints) and C_2 (institutional–governance constraints), via projections π_1 and π_2. Any candidate system X satisfying ϕ_1 and ϕ_2 into the constraint domains factors uniquely through V via the dashed morphism u. A verified aligned system is thus universal among all systems meeting the alignment constraints; the non-existence of u is a categorical certificate of irreducible misalignment.

Part IV

Governance, Deployment, and Societal Stewardship

21 Infrastructure, Maintenance, and Convivial Technology

The mathematical framework of Parts I–III describes alignment as a structural property of functorial mappings between categories. Before turning to the specifically institutional dimensions of alignment governance, it is worth observing that the maintenance of structural integrity over time is a problem that arises not only in AI systems but in any complex sociotechnical infrastructure. Infrastructure studies, software engineering, and ecological management collectively demonstrate that complex systems remain viable through continuous adjustment and repair rather than through episodic redesign [Star, 1999]. The stability of a knowledge system depends on its capacity for structural correction: each maintenance intervention reduces the distortion between the system's current representational state and the semantic domain it is intended to track, which is precisely the dynamic version of the variational alignment condition $\nabla_F \mathcal{E}[F] \to 0$.

Ivan Illich introduced the concept of convivial tools to describe artefacts that enhance the autonomy of their users rather than subordinating individuals to centralized technological systems [Illich, 1973]. Illich's canonical contrast between the bicycle and the automobile is instructive: the bicycle remains repairable, locally maintainable, and powered by the user, while the automobile requires extensive infrastructure and tends to reorganize social patterns around its own demands. In the categorical language of this essay, a convivial AI system is one whose alignment functor $F : \mathcal{H} \to \mathcal{M}$ is interpretable—whose structure is visible enough that the alignment certificate can be read and verified not only by specialists but by affected communities. The interpretability requirements of Part III are the technical conditions for this kind of convivial alignment.

E. F. Schumacher's related concept of intermediate technology emphasized tools that are effective beyond purely manual methods yet simple enough to be maintained locally without dependence on global industrial supply chains [Schumacher, 1973]. The alignment analogue is a verification architecture that is rigorous enough to provide genuine structural guarantees yet modular enough to be audited by institutional actors operating without full access to proprietary model internals. The six-step alignment certificate protocol of Part III is designed with precisely this modularity in mind: each step can in principle be conducted independently, and partial certificates reporting which invariants have been verified (and which residual obstructions remain) are more informative than opaque aggregate safety scores. Maintenance of alignment is therefore not a one-time certification but an ongoing process of iterative structural repair—the continuous reduction of $\mathcal{E}[F]$ as the model, its deployment context, and human normative structures all evolve.

22 The Institutional Meaning of Colimit Preservation

Parts I–III demonstrated that alignment requires the preservation of moral colimits across the representational category $\mathcal{M}$, the RSVP dynamical manifold $\mathcal{X}$, and the action category $\mathcal{A}$. While this analysis provides a precise mathematical account of what alignment is, an additional dimension becomes unavoidable as artificial agents approach real-world deployment: alignment must be embedded within institutional structures that shape, constrain, and verify the behaviour of deployed systems [Amadon et al., 2024, Gabriel, 2020].

From a categorical perspective, institutional governance can be understood as a higher-level fibration that constrains the space of acceptable mappings $G : \mathcal{M} \to \mathcal{A}$. Institutions provide a regulatory category $\mathcal{R}$ of permissible behaviours, morphisms that restrict or modulate internal optimization, and verification mechanisms that observe or audit the mappings within $\mathcal{A}$. The requirement that G preserve normative colimits becomes, at the institutional level, a requirement that all deployed agents satisfy a regulatory functor

$$H : \mathcal{A} \to \mathcal{R}, \tag{19}$$

such that the composite $H \circ G : \mathcal{M} \to \mathcal{R}$ also preserves the relevant colimits. Safe deployment requires that institutional structures make this composite well-defined and stable. Current laws such as the EU AI Act's high-risk classification and reporting thresholds are crude approximations of this idea: they attempt to constrain $\mathcal{A}$ directly through behavioural specifications without reference to the structural conditions that would guarantee colimit preservation. An Alignment Certificate provides a far stricter, structurally grounded version of the same regulatory impulse.

The full picture of institutional alignment can be represented as a commutative cube, showing that normative invariants must be preserved across three orthogonal dimensions simultaneously: the semantic–computational axis, the computational–action axis, and the governance axis.

$$\begin{tikzcd}
\mathcal{H} \arrow[r, "F"] \arrow[d, "J"'] & \mathcal{M} \arrow[d, "G"] & & \mathcal{H}' \arrow[r, "F'"] \arrow[d, "J'"'] & \mathcal{M}' \arrow[d, "G'"] \\
\mathcal{I} \arrow[r, "R"'] & \mathcal{A} & & \mathcal{I}' \arrow[r, "R'"'] & \mathcal{A}'
\end{tikzcd}$$

Figure 6: Commutative cube of alignment across deployment contexts. The left square represents the normative–institutional structure at training time; the right square represents the structure at deployment time. Horizontal morphisms represent semantic encoding (F, F') and action generation (G, G'); vertical morphisms represent institutional constraint (J, J' and R, R'). Alignment requires that all faces commute, so that normative invariants are preserved both across the semantic–action axis and across changes in deployment context. Failures of face commutativity correspond to specific types of misalignment: semantic $(F \neq F')$, behavioural $(G \neq G')$, or governance $(R \neq R')$.

Proposition 22.1 (Structural Alignment Principle)**.** *Let $\mathcal{H}$, $\mathcal{M}$, $\mathcal{A}$, and $\mathcal{I}$ be categories as defined above, with functors $F : \mathcal{H} \to \mathcal{M}$, $G : \mathcal{M} \to \mathcal{A}$, $J : \mathcal{H} \to \mathcal{I}$, and $R : \mathcal{I} \to \mathcal{A}$. The system is institutionally aligned if the square*

$$G \circ F \;\simeq\; R \circ J \tag{20}$$

holds up to natural isomorphism, and if both $G \circ F$ and $R \circ J$ preserve the normative colimits of Definition 1.1. Misalignment arises whenever any face of the commutative square fails, producing structural distortion between normative semantics, computational representations, and real-world actions that is not compensated by the institutional path.

Sketch. If $G \circ F \simeq R \circ J$ and both paths preserve colimits, then for any normative diagram D_N with colimit H_N in $\mathcal{H}$, the action-level image $(G \circ F)(H_N)$ is isomorphic to the institutionally mediated image $(R \circ J)(H_N)$. Both therefore satisfy the colimit-preservation condition of Definition 1.1. If either path fails to preserve colimits, or if the two paths yield non-isomorphic action-level objects, the system exhibits structural distortion along at least one dimension of the alignment cube. □

23 Regulatory Implications of the Representation–Action Gap

A central implication of Parts I–III is that regulatory oversight must focus on $\mathcal{A}$ rather than $\mathcal{M}$. Most existing evaluations—benchmarks of value understanding, moral reasoning tests, preference modelling—only measure representational content. They identify the presence of colimits in $\mathcal{M}$ but not the integrity of their descent into $\mathcal{A}$. This represents a systemic blind spot that the categorical analysis makes precise. Models trained on human data routinely pass moral reasoning benchmarks yet fail under adversarial prompting [Zou et al., 2023]. Reinforcement-learning fine-tuning can overwrite or bypass representational-level moral constraints [Gao et al., 2023]. Capabilities can emerge in $\mathcal{A}$ with little or no precursor signals in $\mathcal{M}$ [Schaeffer et al., 2024]. Regulatory regimes that evaluate systems solely on their semantic fluency risk producing the same category error analyzed in Part I: mistaking representational coherence for alignment.

A growing body of empirical evidence confirms that semantic mastery does not constrain action. Models capable of sophisticated moral reasoning produce harmful outputs under distribution shift [Perez et al., 2022]. Reinforcement- learning agents routinely learn reward-hacking strategies that violate intended goals [Amodei et al., 2016]. Systems trained to follow instructions often fail to generalize norms to new contexts [Shah et al., 2022]. These phenomena exemplify the categorical obstructions analyzed in Parts I and II. Although no existing model implements the full RSVP architecture, several empirical results support its conceptual foundations: interpretability work reveals latent structures that behave like geometric manifolds [Elhage et al., 2022], recurrent models exhibit attractor dynamics that stabilize semantic regions [Yang et al., 2019], and multi-layer reasoning systems show distributed phase transitions reminiscent of RSVP field evolution [Nanda, 2023]. These findings suggest that RSVP dynamics are plausible substrates for representational continuity, though not yet sufficient for alignment without further structural constraints. Policy must therefore incorporate adversarial behavioural audits, stress tests, and formal verification of motivational invariants—the core components developed in Part III.

24 Alignment Certificates as Governance Artefacts

The formal evaluation pipeline developed in Part III naturally extends into a governance artefact: the alignment certificate. A certificate includes explicit identification of normative colimits in $\mathcal{M}$, formal verification of RSVP dynamical stability, adversarial stress-test results, interpretability analyses of G, and proofs or partial proofs of functoriality. Certificates serve two functions. First, they allow institutions to evaluate whether a system meets the structural alignment criteria developed in this essay. Second, they create auditability and accountability: system developers must publicly attest to the preservation of specific universal constructions, not generic assurances about safety or benevolence. This shifts evaluation from textual claims about alignment to categorical, testable claims subject to independent verification.

25 Societal Constraints, Incentive Design, and the Economics of Alignment

Institutions themselves can break the $\mathcal{M} \to \mathcal{A}$ mapping. Economic pressures frequently misalign developer incentives, creating structural conflicts that operate independently of any individual model's architectural properties. Incentives to maximize engagement or revenue can induce misaligned agent behaviour even when $\mathcal{M}$ contains robust normative colimits [Zeng et al., 2023]. Competitive deployment races reduce time for verification and testing [Cotra, 2022]. Organizations may suppress safety findings that slow product cycles [Leike, 2022]. Alignment must therefore include the alignment of institutions as well as agents: governance must enforce constraints that make safe architectures economically viable and unsafe architectures costly. This is not merely a regulatory observation but a structural one—the composition $H \circ G$ can fail at H just as surely as at G, and no amount of architectural soundness in the agent compensates for an institutional functor that does not preserve normative colimits.

26 Multi-Agent Dynamics and Long-Term Stability Under RSVP

As artificial agents become embedded in multi-agent environments, additional obstructions appear. The mapping $G : \mathcal{M} \to \mathcal{A}$ must now commute not only with internal optimization but with external strategic pressures. RSVP provides tools for analyzing multi-agent systems as coupled dynamical fields: agents correspond to intersecting fibres in a global manifold, and social equilibria correspond to stable configurations of field interactions. Stability analysis can detect runaway competitive dynamics, coordination failures, the emergence of adversarial subagents, and the erosion of normative invariants at the population level. These analyses support the design of institutional constraints that preserve moral invariants across agent populations.

27 Toward a Theory of Constitutional AI Under RSVP

A constitutional agent [Bai et al., 2022] can be interpreted categorically as an agent whose action policies are constrained by a set of higher-order invariants. RSVP extends this concept by embedding constitutional constraints into the dynamics of the manifold itself. Rather than applying a rulebook to $\mathcal{A}$ externally, RSVP enforces invariants as dynamical fixed points: trajectories that violate constitutional constraints incur increased action cost under the Lagrangian, making deviation dynamically expensive rather than merely prohibited. This provides a principled method for designing AI systems whose motivational structure is stabilized by universal constructions rather than ad hoc patches or reward functions. The categorical and institutional framework together ensure that these structural guarantees remain meaningful across deployment contexts, updates to the model, and changes in the regulatory environment.

28 Conclusion to Part IV

Alignment must be embedded within governance, institutions, and society. Preservation of normative colimits across $\mathcal{M}$, $\mathcal{X}$, and $\mathcal{A}$ is necessary but not sufficient; the broader environment must enforce structures that support these mappings. The categorical error analyzed in Part I becomes a societal error if institutions equate semantic mastery with motivational safety. Alignment is therefore both a mathematical and a civic responsibility: the categorical and RSVP tools developed in this essay provide a coherent framework for designing agents whose motivations preserve semantic universals, and governance structures must ensure these guarantees are upheld in deployment.

Part V

Deepening the Categorical Program

29 Alignment as Functorial Structure

The preceding parts developed the alignment problem in terms of colimit preservation: a system is aligned if its representational and action categories are linked by a functor G that carries moral colimits to corresponding action-level universals. Part V deepens this program by progressively introducing richer categorical structures—natural transformations, monoidal categories, sheaves, endofunctors, adjunctions, and variational principles—each of which captures a distinct structural dimension of the alignment problem not fully addressed by the colimit-preservation requirement alone.

The central claim of Part V is that alignment is fundamentally a structural relationship between domains of representation, and that this relationship admits a layered mathematical treatment in which each successive formalism resolves ambiguities and failure modes that the previous layer could not fully address. Rather than treating alignment as a collection of constraints imposed on individual systems, we model it as a structure-preserving correspondence between two categories, and we show that progressively richer structures are needed to capture corrigibility, compositionality, contextual consistency, informational distortion, recursive stability, and multi-agent coherence.

Let $\mathcal{H}$ denote the category of human semantic structures, whose objects represent coherent semantic states such as beliefs, intentions, or conceptual schemas, and whose morphisms represent transformations of meaning including inference, revision, or contextual reinterpretation. Let $\mathcal{M}$ denote the category of machine representations, whose objects correspond to internal representational states of a computational system and whose morphisms correspond to state transitions induced by computation. Alignment may then be expressed as the existence of a functor

$$F : \mathcal{H} \to \mathcal{M} \tag{21}$$

that preserves the relevant structural relations between semantic states. For any pair of composable morphisms $f : A \to B$ and $g : B \to C$ in $\mathcal{H}$, the functor must satisfy

$$F(g \circ f) \;=\; F(g) \circ F(f), \tag{22}$$

together with preservation of identity morphisms.

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[d, "F"'] & B \arrow[d, "F"] \\
F(A) \arrow[r, "F(f)"'] & F(B)
\end{tikzcd}$$

Figure 7: Functorial alignment. Semantic transformations in the human category are mapped to corresponding transformations in the machine representation category. Alignment requires the diagram to commute for all morphisms f in $\mathcal{H}$.

Under this interpretation, alignment failures correspond to violations of structure preservation. If semantic transformations that are coherent within $\mathcal{H}$ are mapped to incoherent or divergent transformations in $\mathcal{M}$, then the functor fails to preserve composition, and the resulting mismatch manifests operationally as unintended system behaviour. This framing naturally explains many alignment failures: they correspond to functors that fail to preserve limits, colimits, or other structural invariants, not to failures of any individual rule or specification.

30 Natural Transformations and Corrigibility

Real systems rarely operate under a single interpretation of their objective structure. Instead, multiple semantic mappings coexist: a machine may internally represent a goal structure that differs from the normative interpretation imposed by its designers, and updating one interpretation to align with another is precisely the problem of corrigibility. Category theory provides a precise language for describing this situation through natural transformations.

Let $F, G : \mathcal{H} \to \mathcal{M}$ be two functors representing distinct interpretations of the mapping between human semantics and machine representations. A natural transformation $\eta : F \Rightarrow G$ consists of a family of morphisms $\eta_A : F(A) \to G(A)$ for each object A in $\mathcal{H}$ such that the following diagram commutes for every morphism $f : A \to B$ in $\mathcal{H}$:

$$\begin{tikzcd}
F(A) \arrow[r, "F(f)"] \arrow[d, "\eta_A"'] & F(B) \arrow[d, "\eta_B"] \\
G(A) \arrow[r, "G(f)"'] & G(B)
\end{tikzcd} \tag{23}$$

This commutativity condition expresses a key structural property of corrigibility: the adjustment from one interpretation to another preserves the structure of semantic transformations throughout. Corrigibility is thus not merely the ability to modify a system's parameters; it is the existence of coherent transformations between interpretive functors that maintain structural consistency across the system's entire reasoning process. A system is corrigible with respect to a normative correction precisely when there exists a natural transformation from its current interpretive functor to the corrected one, ensuring that no semantic relationship is distorted by the update.

$$\begin{tikzcd}
F(A) \arrow[r, "F(f)"] \arrow[d, "\eta_A"'] & F(B) \arrow[d, "\eta_B"] \\
G(A) \arrow[r, "G(f)"'] & G(B)
\end{tikzcd}$$

Figure 8: Corrigibility as a natural transformation. The commutativity condition ensures that adjustment from the machine's interpretive functor F to the normative functor G preserves the structure of semantic transformations. Each morphism η_A is a local corrigibility witness at the object A.

This formulation has a further consequence: the collection of all natural transformations between two functors $F, G : \mathcal{H} \to \mathcal{M}$ forms a set $\mathrm{Nat}(F, G)$, and in suitable settings this set itself carries further categorical structure. The existence and properties of natural transformations between the machine's current interpretive functor and the normatively intended one can therefore be studied as an algebraic object in its own right, providing a formal basis for questions such as: how many distinct corrigibility paths exist, which are minimal, and which preserve additional invariants such as safety constraints.

31 Monoidal Composition and Modular Alignment

Modern computational systems are rarely monolithic. They are pipelines composed of many interacting subsystems whose outputs serve as inputs to others, and alignment must be analyzed not only for individual components but for their compositions. Category theory already provides a language for describing compositional systems through symmetric monoidal categories, and Part II's RSVP construction extends naturally to this setting.

Let $(\mathcal{C}, \otimes, I)$ denote a symmetric monoidal category in which objects represent computational modules and the tensor product $\otimes$ represents parallel composition. Suppose each subsystem admits a functorial alignment mapping $F_i : \mathcal{H}_i \to \mathcal{M}_i$. The monoidal structure allows construction of a combined system:

$$F_1 \otimes F_2 : \mathcal{H}_1 \otimes \mathcal{H}_2 \to \mathcal{M}_1 \otimes \mathcal{M}_2. \tag{24}$$

Alignment is compositional if the tensor product preserves the structural properties required for alignment. A key question is whether colimit preservation is closed under this operation: if F_1 and F_2 each individually preserve the relevant colimits in $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively, does $F_1 \otimes F_2$ preserve the colimits of the combined diagram?

$$\begin{tikzcd}
H_1 \otimes H_2 \arrow[r, "f_1 \otimes f_2"] \arrow[d, "F_1 \otimes F_2"'] & H'_1 \otimes H'_2 \arrow[d, "F_1 \otimes F_2"] \\
M_1 \otimes M_2 \arrow[r, "F_1(f_1) \otimes F_2(f_2)"'] & M'_1 \otimes M'_2
\end{tikzcd}$$

Figure 9: Compositional alignment under monoidal structure. Alignment should be preserved when independently aligned subsystems are combined through the tensor product. Failures of this diagram to commute correspond to emergence of misalignment at the compositional level.

The answer is not automatically yes. Failures of compositional alignment arise when interactions between subsystems introduce morphisms that violate the preservation conditions of the underlying functors: information flow from one module to another can create correlations that break the independence assumptions implicit in the colimit structure of each component, and these correlations can manifest as new obstructions absent from any individual subsystem's analysis. This observation explains why individually well-behaved components can produce misaligned behaviour when integrated into larger architectures, and it implies that compositional verification—not merely component-level verification—is a necessary condition for safe deployment of modular systems.

32 Sheaf-Theoretic Alignment and Contextual Consistency

Many alignment problems arise specifically from context sensitivity: a model behaves acceptably in one domain but fails when the same reasoning is extended elsewhere. The failure is not a failure of any single local behaviour but a failure of local behaviours to assemble into a globally coherent policy. Sheaf theory provides a precise mathematical language for formalizing local consistency and the conditions under which it extends to global coherence.

Consider a topological space X representing the space of semantic contexts. For each open set $U \subseteq X$, let $\mathcal{F}(U)$ denote the set of machine behaviours consistent with the semantic constraints of context U. The assignment $U \mapsto \mathcal{F}(U)$ defines a presheaf over the context space, equipped with restriction maps $\mathcal{F}(V) \to \mathcal{F}(U)$ for each inclusion $U \subseteq V$. Alignment requires this presheaf to be a sheaf—that is, to satisfy the gluing condition: whenever a family of local sections $s_i \in \mathcal{F}(U_i)$ agrees on pairwise overlaps $U_i \cap U_j$, there exists a unique global section

$$s \in \mathcal{F}\Big(\bigcup_i U_i\Big) \tag{25}$$

whose restrictions recover all the local sections.

$$\begin{tikzcd}
& \mathcal{F}(U_1 \cup U_2) \arrow[dl] \arrow[dr] & \\
\mathcal{F}(U_1) \arrow[dr] & & \mathcal{F}(U_2) \arrow[dl] \\
& \mathcal{F}(U_1 \cap U_2) &
\end{tikzcd}$$

Figure 10: Sheaf gluing condition. Local semantic behaviours on contexts U_1 and U_2 must agree on their overlap in order to form a globally aligned interpretation. Obstruction to gluing corresponds to contextual misalignment.

When this gluing condition fails, the system exhibits contextual misalignment: behaviours that appear coherent within individual domains fail to combine into a globally consistent semantic structure. The obstruction to gluing is precisely the first sheaf cohomology class $H^1(X, \mathcal{F})$: this class is non-trivial when and only when compatible local sections cannot be assembled into a global one. Sheaf-theoretic alignment thus provides an exact measure of contextual inconsistency, and the verification condition $H^1(\mathcal{U}, \mathcal{A}) = 0$ derived in Part III now receives a precise interpretation: it is the sheaf condition guaranteeing that the agent's locally moral behaviour assembles into a globally moral policy.

33 Information Flow and Structural Distortion

Mappings between representational systems inevitably involve information compression and transformation, and these transformations may introduce distortions that reduce the fidelity of the alignment functor. The categorical framework developed so far treats alignment in purely structural terms—functors, limits, natural transformations, and adjunctions—but many practical systems require a quantitative measure of how much structure is lost when representations are translated between domains. Information theory provides a natural language for expressing this idea.

Let $F : \mathcal{H} \to \mathcal{M}$ be a functor representing the translation from human semantic structures to machine representations. Consider a probability distribution P_H over objects of $\mathcal{H}$, representing the frequency with which different semantic states are encountered. The functor F induces a corresponding distribution over $\mathcal{M}$, denoted $P_M = F_* P_H$. The divergence between these distributions may be measured using the Kullback–Leibler divergence:

$$D_{\mathrm{KL}}(P_H \,\|\, P_M) \;=\; \sum_{H} P_H(H) \log \frac{P_H(H)}{P_M(F(H))}, \tag{26}$$

which quantifies the structural distortion introduced by the mapping. More generally, morphisms in $\mathcal{M}$ may be interpreted as channels that transform probability distributions over representational states. Let $f : M_1 \to M_2$ be such a morphism; the entropy change associated with the induced transformation of the distribution is $\Delta H = H(P_{M_2}) - H(P_{M_1})$. Positive entropy change corresponds to loss of structural information, while negative entropy change indicates the introduction of additional constraints.

Within this framework, alignment may be interpreted as the minimization of information-theoretic distortion across the morphisms that connect semantic and computational domains. This perspective suggests that alignment is not a binary property but a continuous quantity that can be optimized subject to computational and architectural constraints, admitting a notion of alignment budget: the maximum structural distortion that a given deployment context can tolerate while remaining within acceptable safety bounds.

34 Alignment as a Fixed Point of Recursive Endofunctors

Intelligent systems are rarely static. Most modern architectures update their internal representations continuously through learning, inference, or iterative refinement, and alignment must remain stable under these updates. The categorical language for describing recursive dynamics is provided by endofunctors and their fixed points.

Let $\mathcal{M}$ denote the category of machine representational states. A learning or update procedure can be modeled as an endofunctor

$$T : \mathcal{M} \to \mathcal{M}. \tag{27}$$

Each application of T corresponds to a single update step in the system's internal model. If alignment is defined by a functor $F : \mathcal{H} \to \mathcal{M}$, the stability of alignment depends on how the update operator T interacts with the image of F. Alignment is preserved precisely when:

$$T(F(H)) \;\cong\; F(H) \tag{28}$$

for the relevant semantic states H—that is, aligned states form a fixed point or invariant subcategory under the update dynamics. More generally, one considers fixed points of the endofunctor T itself: an object M is a fixed point when $T(M) \cong M$.

$$\begin{tikzcd}
M \arrow[r, "T"] \arrow[dr, equal] & T(M) \arrow[d, "\sim"] \\
& M
\end{tikzcd}$$

Figure 11: Fixed-point stability of the update endofunctor T. Alignment is preserved when aligned states remain invariant under the system's learning dynamics. The isomorphism $T(M) \cong M$ expresses dynamical stability of the aligned representational state.

Such fixed points represent stable internal representations under the system's learning dynamics. If these fixed points correspond to images of human semantic structures under the functor F, then alignment remains stable under recursive self-improvement. Conversely, alignment drift can be interpreted as a bifurcation in which the attractors of the update dynamics move outside the image of the alignment functor: the system's internal model converges to a fixed point of T that does not correspond to any legitimate semantic state in the image of F.

The categorical theory of fixed points of endofunctors has been extensively developed through the theory of initial algebras and final coalgebras [Adamek & Rosicky, 2005]: an initial T-algebra is the least fixed point of T, and a final T-coalgebra is the greatest. Alignment stability under recursive self-improvement corresponds, in this framework, to the condition that the image of F is contained in the initial algebra of the update endofunctor—the minimal stable structure that all update paths must traverse. This condition is stronger than mere attractor stability and provides a framework for reasoning about systems that modify their own learning procedures.

35 Higher Categorical Structures and Multi-Agent Alignment

Real-world AI systems do not exist in isolation. They interact with other systems, with human users, and with institutional structures, and these interactions introduce higher-order relationships that cannot be fully captured within ordinary categories. To describe such systems it is natural to extend the framework to higher categories, in which morphisms themselves admit morphisms between them.

Let $\mathcal{C}$ be a 2-category whose objects represent agents or cognitive systems, whose 1-morphisms represent communication channels or information transformations between agents, and whose 2-morphisms represent transformations between these interaction processes. Suppose each agent A_i admits a semantic functor $F_i : \mathcal{H}_i \to \mathcal{M}_i$. Interactions between agents induce additional functors between the categories $\mathcal{M}_i$: communication between two systems may be modeled as a functor $C_{ij} : \mathcal{M}_i \to \mathcal{M}_j$. Alignment across the multi-agent system requires that the composite structures formed by these functors satisfy higher-order coherence conditions. Semantic meaning must remain consistent when information flows through chains of agents and transformations, and this requirement is expressed through commutative diagrams of functors and natural transformations within the 2-category.

Misalignment in multi-agent systems often arises not from individual components but from incoherence among these higher-order transformations. Information that remains semantically consistent within one subsystem may become distorted when transported through a sequence of interacting systems, and this distortion may not be detectable at the level of any individual agent. The higher categorical perspective therefore shifts the focus of alignment research from individual models to the structural coherence of entire interaction networks, and it implies that alignment certification for multi-agent systems must verify the commutativity of diagrams involving 1-morphisms and 2-morphisms jointly—a substantially more demanding requirement than certification for isolated agents.

36 Limits, Colimits, and Semantic Merging

The notions of limit and colimit, introduced in Part I as the formal description of how human normative language generates canonical representational attractors, admit a richer analysis in the context of Part V's extended categorical program. Many alignment problems arise when information from multiple sources must be combined into a single coherent representation: systems integrating heterogeneous data, merging knowledge from different models, or reconciling competing interpretations of a situation must perform operations that are precisely captured by these categorical constructions.

Let $\mathcal{C}$ denote a category of semantic structures and let $D : J \to \mathcal{C}$ be a diagram representing a collection of related semantic objects indexed by a small category J. A limit of the diagram D is an object L equipped with morphisms $\pi_j : L \to D(j)$ for each object j in J such that every compatibility relation in the diagram is respected, and such that L is universal with respect to this property: any other compatible cone factors uniquely through L. In the context of alignment, limits correspond to the construction of a representation that simultaneously satisfies multiple semantic constraints. A system integrating ethical rules, contextual knowledge, and observed data may be interpreted as computing a limit in an appropriate semantic category, and failures of alignment arise when the system constructs a suboptimal cone—one that satisfies some constraints while violating others—rather than the universal one.

Dually, a colimit of the diagram D is an object C equipped with morphisms $\iota_j : D(j) \to C$ such that any other object receiving compatible morphisms from the diagram factors uniquely through C. Colimits represent semantic aggregation: when multiple sources of information contribute partial knowledge about a domain, their colimit produces a unified structure that incorporates all contributions while identifying shared components. Misalignment frequently arises when the system constructs colimits that identify structures which should remain distinct, collapsing important semantic distinctions, or when limits impose constraints that eliminate relevant distinctions. Alignment requires ensuring that the categorical constructions used to merge semantic structures preserve the intended interpretive relations across all scales of the system's architecture.

37 Adjunctions and Interpretation Layers

The functorial formulation of alignment describes a one-directional mapping from human semantic structures to machine representations. However, interpretation between these domains is typically bidirectional: abstraction maps machine states to human interpretation, while implementation maps human concepts to machine states. Category theory captures this bidirectional relationship through adjoint functors, which provide the most refined structure relating two categories connected by a pair of mappings.

Let $\mathcal{H}$ denote the category of human semantic structures and $\mathcal{M}$ the category of machine representations, and suppose there exist functors $F : \mathcal{H} \to \mathcal{M}$ and $G : \mathcal{M} \to \mathcal{H}$ translating between the two domains. The pair (F, G) forms an adjunction $F \dashv G$ when there exists a natural isomorphism

$$\mathrm{Hom}_{\mathcal{M}}(F(H), M) \;\cong\; \mathrm{Hom}_{\mathcal{H}}(H, G(M)) \tag{29}$$

for all objects $H \in \mathcal{H}$ and $M \in \mathcal{M}$.

$$\begin{tikzcd}
\mathcal{H} \arrow[r, shift left=2, "F"] \arrow[r, phantom, "\perp"] & \mathcal{M} \arrow[l, shift left=2, "G"]
\end{tikzcd}$$

Figure 12: Adjunction between human semantic structures and machine representations. The functor F encodes semantics into machine states while G interprets machine states semantically. The adjunction condition expresses a tight compatibility between encoding and interpretation.

Associated with the adjunction are the unit and counit morphisms:

$$\eta : \mathrm{Id}_{\mathcal{H}} \Rightarrow G \circ F, \qquad \varepsilon : F \circ G \Rightarrow \mathrm{Id}_{\mathcal{M}}. \tag{30}$$

These transformations measure the fidelity of translation between domains. The unit η describes how well semantic concepts survive encoding into machine representations and subsequent interpretation: the natural transformation $\eta_H : H \to G(F(H))$ witnesses the distortion introduced by the round-trip $H \mapsto F(H) \mapsto G(F(H))$. The counit ε describes how faithfully machine representations implement semantic structures: the transformation $\varepsilon_M : F(G(M)) \to M$ witnesses the distortion introduced by the reverse round-trip. Alignment corresponds to adjunctions in which these distortions remain bounded: when the unit and counit approach isomorphisms, the interpretive loop between human meaning and machine computation becomes structurally stable, and the system can be said to have achieved a tight coupling between its representational and interpretive layers.

38 Alignment as Diagrammatic Constraint Satisfaction

Category theory provides a powerful diagrammatic language for expressing structural relationships between objects and morphisms. These diagrams can be interpreted as systems of constraints that representations must satisfy, and the alignment problem can be reformulated as the problem of constructing representations and update operators that preserve commutativity across a specified family of diagrams.

Let $\mathcal{C}$ be a category whose objects represent semantic or computational states and whose morphisms represent transformations between them. A diagram $D : J \to \mathcal{C}$ specifies a network of relationships indexed by the small category J. A diagram is said to commute when all paths between any two objects yield the same composite morphism. Consider a semantic transformation $f : H_1 \to H_2$ within the human semantic category and its machine counterpart $F(f) : F(H_1) \to F(H_2)$ under the alignment functor F. If a system update operator T acts on machine states, alignment requires the following diagram to commute:

$$\begin{tikzcd}
F(H_1) \arrow[r, "F(f)"] \arrow[d, "T"'] & F(H_2) \arrow[d, "T"] \\
T(F(H_1)) \arrow[r, "T(F(f))"'] & T(F(H_2))
\end{tikzcd} \tag{31}$$

When such diagrams fail to commute, the system introduces structural distortions into the relationship between semantic and computational transformations. The alignment problem may therefore be formulated as a diagrammatic constraint satisfaction problem: the objective is to construct representations and update operators that preserve commutativity across all diagrams in a specified family. Rather than specifying individual rules or prohibitions, one specifies families of diagrams whose commutativity defines acceptable system behaviour. This formulation connects alignment research with the tradition of constraint-based reasoning in mathematics and theoretical computer science, and it clarifies the sense in which alignment is a structural property of a system rather than a property of any individual output.

39 A Variational Principle for Alignment

The categorical framework developed across Parts I–V describes alignment as the preservation of structural relations between semantic and computational domains. In practice, however, perfect preservation is rarely achievable under the constraints of computation, representation, and noise. It is therefore natural to treat alignment not as an exact categorical property but as the solution of a variational problem: alignment is defined as the minimization of a structural distortion functional over the space of admissible mappings between semantic and computational systems.

Let $\mathrm{Fun}(\mathcal{H}, \mathcal{M})$ denote the space of functors from the human semantic category to the machine representational category. For each such functor F we define a structural distortion functional $\mathcal{E}[F]$ measuring the degree to which F fails to preserve the structural relations of the semantic domain. This functional has three natural components.

A first component measures functorial distortion. Given morphisms $f : H_1 \to H_2$ and $g : H_2 \to H_3$ in $\mathcal{H}$, a perfectly structure-preserving mapping satisfies $F(g \circ f) = F(g) \circ F(f)$, and deviations from this condition contribute an energy $E_{\text{functor}}(F)$ proportional to the magnitude of the violation summed over all composable morphism pairs.

A second component measures information-theoretic distortion: if P_H denotes a distribution over semantic states and $F_* P_H$ the induced distribution over machine states, the divergence $D_{\mathrm{KL}}(P_H \| F_* P_H)$ quantifies the loss of semantic information under the representation mapping, defining the energy $E_{\text{information}}(F)$.

A third component measures contextual inconsistency. When semantic structures are organized as a sheaf over a context space X, alignment requires that compatible local sections glue into a global section, and the failure of gluing conditions contributes energy $E_{\text{context}}(F)$ proportional to the magnitude of the obstruction class in $H^1(X, \mathcal{F})$.

Combining these components yields the structural distortion functional:

$$\mathcal{E}[F] \;=\; \lambda_1 E_{\text{functor}}(F) + \lambda_2 E_{\text{information}}(F) + \lambda_3 E_{\text{context}}(F), \tag{32}$$

where the coefficients $\lambda_i > 0$ weight the relative importance of the different sources of distortion. Alignment may then be defined as the solution of the variational problem:

$$F^* \;=\; \underset{F \in \mathrm{Fun}(\mathcal{H}, \mathcal{M})}{\arg\min} \;\mathcal{E}[F]. \tag{33}$$

This formulation connects the categorical description of alignment with optimization principles widely used in machine learning and statistical physics. Learning procedures can be interpreted as iterative processes that attempt to reduce the distortion energy by adjusting the mapping between semantic and computational domains, and stable aligned systems correspond to low-energy configurations of the structural distortion functional.

Proposition 39.1 (Variational Alignment)**.** *Let $\mathcal{E}[F]$ be the structural distortion functional defined above. Among all functors $F : \mathcal{H} \to \mathcal{M}$, the global minima of $\mathcal{E}$ are characterized by three jointly necessary conditions: $E_{\text{functor}}(F) = 0$ (exact colimit preservation),*

$E_{\text{information}}(F) = 0$ *(vanishing KL divergence between the source distribution and its image), and* $E_{\text{context}}(F) = 0$ *(vanishing sheaf obstruction class). In practice, real systems achieve only approximate minima: alignment corresponds to* local *or* approximate *minima of* $\mathcal{E}$*, and the degree of alignment is measured by the magnitude of the residual distortion components at the achieved minimum. The variational alignment certificate of Part III should therefore be understood as a protocol for bounding the residual distortion* $\mathcal{E}[F^*]$ *at the system's operative mapping, not as a protocol for verifying exact zero-distortion optimality.*

The variational perspective provides a unifying interpretation of the preceding sections. Functors describe candidate mappings between domains. Natural transformations describe coherent adjustments between such mappings. Limits and sheaf conditions impose structural constraints on admissible representations. Entropy-based measures quantify the loss of semantic information. The alignment problem then emerges as the task of minimizing the structural distortion induced by these transformations, and the certificate standard developed in Part III can be re-read as a protocol for verifying proximity to the variational optimum.

40 Proxy Optimization and Signaling Collapse

Many sociotechnical systems rely on observable signals as proxies for underlying semantic qualities. These signals are intended to correlate with normative structures but remain only indirect measurements of them. Benchmarks, reward functions, safety scores, and compliance metrics all occupy this role: each is a measurable quantity in an observable proxy space $\mathcal{P}$ chosen to approximate some harder-to-measure normative property in $\mathcal{H}$.

Let $S : \mathcal{H} \to \mathcal{P}$ be the signal extraction functor mapping normative structures to a proxy space of measurable indicators. The intended chain is:

$$\mathcal{H} \xrightarrow{\quad S \quad} \mathcal{P} \xrightarrow{\quad T \quad} \mathcal{A} \tag{34}$$

where $T : \mathcal{P} \to \mathcal{A}$ translates proxy scores into actions. For this chain to be aligned, the composite $T \circ S$ must preserve the normative colimits of $\mathcal{H}$. But S is generically *not* colimit-preserving: it projects out information, collapses distinctions, and introduces equivalence classes that do not respect the structure of $\mathcal{H}$. Two normative states that are categorically distinct—because they imply different action-level obligations—may map to the same proxy score. Once optimization pressure is applied within $\mathcal{P}$, the system dynamics decouple from $\mathcal{H}$ entirely. The proxy space becomes self-referential: actions optimize the observable signal rather than the normative structure the signal was designed to approximate.

This phenomenon is a precise categorical formulation of Goodhart's Law [Manheim &

Garrabrant, 2018]: when a measure becomes a target, it ceases to be a good measure. The failure is structural. It does not arise from bad intentions or insufficient optimization power; it arises because S cannot be colimit-preserving in general, and optimization in $\mathcal{P}$ therefore cannot preserve normative invariants in $\mathcal{H}$. The resulting system remains coherent and even high-performing with respect to the proxy domain while being systematically misaligned relative to the normative structures that motivated the proxy in the first place.

This observation has direct implications for alignment evaluation. Any evaluation protocol that measures only proxy quantities—benchmark performance, RLHF reward scores, red-team pass rates—faces the same structural limitation. The Alignment Certificate of Part III is designed precisely to avoid this collapse: rather than measuring proxies for alignment, it requires direct structural evidence that the composite functor $G \circ F : \mathcal{H} \to \mathcal{A}$ preserves normative colimits. The distinction between structural certification and proxy evaluation is therefore not a matter of degree but a categorical difference in what kind of evidence is being produced.

41 Alignment as a Fixed-Point Condition

The preceding analysis has treated the alignment mappings $F : \mathcal{H} \to \mathcal{M}$ and $G : \mathcal{M} \to \mathcal{A}$ as if they were static. Real intelligent systems are not static: they update their internal representations through continued training, in-context learning, policy refinement, and autonomous modification. Alignment must therefore be understood not merely as a property of an initial mapping but as a dynamical condition that must be maintained across the system's full update trajectory.

Let $U : \mathcal{M} \to \mathcal{M}$ denote the update operator governing changes in the system's internal representation space. After an update step the full behavioral chain becomes:

$$\mathcal{H} \xrightarrow{\quad F \quad} \mathcal{M} \xrightarrow{\quad U \quad} \mathcal{M} \xrightarrow{\quad G \quad} \mathcal{A} \tag{35}$$

A system is *dynamically aligned* if $G \circ U \circ F$ preserves the same normative colimits as the original $G \circ F$. This is the requirement that U belong to the class of *alignment-preserving update operators*: those endofunctors of $\mathcal{M}$ that commute with the colimit-preservation property of the composed chain. Not every update operator satisfies this condition. In-weight learning may shift the geometry of $\mathcal{M}$ in ways that break previously stable fibrations; reward-driven policy modification may push the action category $\mathcal{A}$ toward new attractors that no longer coincide with normative basins; recursive self-improvement may iterate U many times, compounding small structural distortions into large misalignments.

In categorical terms, alignment under updates corresponds to the fixed-point condition:

$$\big[G \circ U^n \circ F \text{ is colimit-preserving}\big] \quad \text{for all } n \geq 0, \tag{36}$$

which is the requirement that the sequence $\{G \circ U^n \circ F\}_{n \geq 0}$ lies entirely within the subspace of colimit-preserving functors $\mathcal{H} \to \mathcal{A}$. A sufficient condition for this is that U itself be a colimit-preserving endofunctor of $\mathcal{M}$: if U preserves the semantic structure of $\mathcal{M}$, then iterating U cannot destroy the colimit structure that F placed there. This is not automatically satisfied by gradient descent, parameter averaging, or any other standard update rule, and designing update operators that provably lie within this class is a non-trivial alignment engineering problem.

The fixed-point formulation also clarifies the danger in systems capable of recursive self-improvement: if U is itself learned or generated by the system, there is no guarantee that successive iterates remain within the alignment-preserving class, and small systematic deviations compound. Ensuring alignment of self-improving systems therefore requires constraints not only on the initial mappings but on the *space of admissible update operators*, which is a higher-order version of the alignment problem.

42 Compositional Alignment Across Scales

The categorical framework developed throughout this paper describes alignment as a property of individual systems. But real-world deployment involves collections of agents operating within shared institutional environments, and the compositional structure of alignment across such collections presents challenges that do not arise at the individual level.

Consider a collection of n agents with representation categories $\mathcal{M}_1, \ldots, \mathcal{M}_n$ and action categories $\mathcal{A}_1, \ldots, \mathcal{A}_n$. Each agent participates in a behavioral chain:

$$\mathcal{H} \xrightarrow{F_i} \mathcal{M}_i \xrightarrow{G_i} \mathcal{A}_i, \tag{37}$$

and individual alignment requires each $G_i \circ F_i$ to be colimit-preserving. At the collective level the agents' actions compose through a world-transition functor:

$$\Gamma : \bigotimes_{i=1}^{n} \mathcal{A}_i \ \to \ \mathcal{W}, \tag{38}$$

where $\mathcal{W}$ is the category of world states and $\bigotimes$ denotes the monoidal product of action categories. System-level alignment requires the composite

$$\Gamma \circ \big(G_1 \circ F_1 \otimes \cdots \otimes G_n \circ F_n\big) : \mathcal{H}^{\otimes n} \to \mathcal{W} \tag{39}$$

to preserve normative invariants as well. This is a strictly stronger condition than individual alignment: even if every $G_i \circ F_i$ is colimit-preserving, the monoidal functor Γ may fail to preserve the colimits of the product category. Individual alignment is necessary but not sufficient for collective alignment.

This gap between individual and collective alignment corresponds to well-known phenomena in complex systems: locally rational agents can generate globally irrational outcomes through the interaction of their individually optimal policies. What the categorical framework adds is a precise structural account of *why* this happens and what would need to be true for it not to. The necessary condition is that Γ be colimit-preserving as a monoidal functor, which in practical terms requires coordination mechanisms—shared representations, communication protocols, institutional constraints, or joint policy structures— that couple the individual chains into a coherent collective one. Absent such mechanisms, the alignment of individual agents provides no guarantee about the alignment of the system they collectively constitute.

The implication for governance is direct: institutional structures must be designed not merely to ensure that individual agents pass alignment certification, but to ensure that the aggregation functor Γ preserves the normative colimits of the product category. This is the multi-agent analogue of the Structural Alignment Principle (Proposition 22.1), and it suggests that collective alignment is fundamentally an institutional design problem as much as an individual engineering one.

43 Structural Alignment Theorem

The preceding sections have developed the alignment problem across four dimensions: the representation–motivation gap, the constructive RSVP architecture, proxy optimization failure, and the fixed-point condition under updates. These threads can now be drawn together into a single theorem that states the conditions under which alignment is preserved across structural composition and dynamic evolution simultaneously.

Theorem 43.1 (Structural Alignment)**.** *Let $\mathcal{H}$ be the category of human normative structures (Definition 1.2), $\mathcal{M}$ the category of system representations, $\mathcal{A}$ the category of actions, and $\mathcal{W}$ the category of world states. Let*

$$F : \mathcal{H} \to \mathcal{M}, \quad G : \mathcal{M} \to \mathcal{A}, \quad C : \mathcal{A} \to \mathcal{W}$$

be functors, and let $U : \mathcal{M} \to \mathcal{M}$ be an update operator. The system is structurally and dynamically aligned *if and only if:*
(i) $G \circ F : \mathcal{H} \to \mathcal{A}$ is colimit-preserving for all normative diagrams D_N;
(ii) U is colimit-preserving as an endofunctor of $\mathcal{M}$, so that $G \circ U^n \circ F$ remains colimit-preserving for all $n \geq 0$;
(iii) the obstruction class $\omega(F, G) \in H^1(\mathcal{U}, \mathcal{A})$ vanishes, ensuring that locally aligned behaviours assemble into a globally coherent policy;
(iv) for any institutional constraint functor $J : \mathcal{H} \to \mathcal{I}$ and $R : \mathcal{I} \to \mathcal{A}$, the diagram $G \circ F \simeq R \circ J$ commutes up to natural isomorphism.

Whenever any of conditions (i)–(iv) fails, structural distortion is introduced between $\mathcal{H}$ and $\mathcal{W}$: the system produces observable outcomes that violate the normative invariants of $\mathcal{H}$.

Sketch. Condition (i) is the colimit-preservation requirement of Definition 1.1. Condition (ii) ensures that the update dynamics preserve the structural invariants placed by F: if U is colimit-preserving and F is colimit-preserving, then $U \circ F$ is colimit-preserving by composition of cocontinuous functors [Mac Lane, 1998], and the result extends to all iterates by induction. Condition (iii) is the sheaf-coherence requirement: if $\omega(F, G) \neq 0$, there exist overlapping contexts on which locally consistent behaviours cannot be assembled into a global policy, producing contextual misalignment that is not detectable from any single context alone. Condition (iv) is the institutional commutativity requirement of Proposition 22.1: if the direct path and the institutional path through $\mathcal{I}$ disagree, the governance structure fails to enforce the intended normative constraints. Each failure mode is therefore both structurally distinct and irreducible: no combination of the remaining three conditions can compensate for the failure of any one. □

The theorem unifies the major positive results of the paper. Definition 1.1 provides condition (i); the RSVP architecture of Part II provides a dynamical substrate for condition (ii) under the regularity conditions of Theorem 13.1; the sheaf-cohomological diagnostics of Part III and Appendix G provide a test for condition (iii); and the Structural Alignment Principle of Proposition 22.1 provides condition (iv). The Alignment Certificate of Part III can therefore be read as an operational checklist for the four conditions of Theorem 43.1, and alignment research as the project of engineering systems and institutions that jointly satisfy them.

Concluding Reflections: Maintenance, Perception, and Structural Stewardship

The variational formulation of alignment arrived at in Part V bears a structural resemblance to principles that appear across several domains outside alignment research, and tracing these resemblances serves not to extend the formal machinery but to locate the alignment problem within a broader intellectual landscape. Three such resemblances are worth identifying, not as analogies but as instances of a common underlying pattern: systems that remain viable over time do so by continuously correcting model–environment mismatch rather than by achieving perfect representation.

The first instance appears in the theory of predictive processing. Anil Seth and others have characterized perception as a form of "controlled hallucination" in which the brain functions as a generative model that continuously predicts the causes of its sensory inputs and adjusts its internal model to minimize discrepancies between prediction and observation [Seth, 2021, Friston, 2010]. Formally, predictive processing is described using the principle of variational free-energy minimization. Let s denote sensory signals and z latent variables representing hidden environmental causes. The brain maintains a generative model $p(s, z)$ and an approximate posterior distribution $q(z)$ over latent causes. The variational free energy is:

$$F[q] \;=\; D_{\mathrm{KL}}\big(q(z) \,\|\, p(z \mid s)\big) - \log p(s). \tag{40}$$

Minimizing this quantity drives the internal model toward states that better explain incoming sensory data while maintaining internal consistency. This is a special case of the alignment functional $\mathcal{E}[F]$ defined in Section V.9: the generative model $p(s, z)$ is the alignment functor, the divergence term is the information-theoretic distortion component, and the log-evidence term regularizes the prior structure of the model. Aligned systems and perceiving organisms are, from this viewpoint, instances of the same variational structure: systems that minimize distortion between their internal representations and the constraints imposed by their environment.

The second instance concerns maintenance and repair as epistemic practices. Infrastructure studies, software engineering, and ecological management collectively demonstrate that complex systems remain viable through continuous adjustment rather than episodic redesign [Star, 1999]. In this perspective, alignment between a model and the world is not a static property but an ongoing maintenance process. Representational structures drift as environments change, new information appears, and previous assumptions become obsolete. Repairing these structures requires iterative refinement whose structure is precisely that of gradient descent on the distortion functional $\mathcal{E}$: each maintenance step reduces the distortion between the system's current representational state and the semantic category it is intended to track. The stability of a knowledge system therefore depends on its capacity for continuous structural correction, which is the dynamic version of the fixed-point condition $T(F(H)) \cong F(H)$ introduced in Section V.7.

The third instance concerns the design of technology for maintenance and resilience. Ivan Illich's concept of convivial tools—tools that enhance the autonomy of their users rather than subordinating individuals to centralized systems [Illich, 1973]—expresses a design principle with direct categorical content. A convivial tool is one whose internal structure is interpretable and repairable by those who use it: its implementation is transparent enough that the users can modify, adapt, and maintain it without expert intermediation. In the categorical language of this essay, a convivial AI system is one whose alignment functor $F : \mathcal{H} \to \mathcal{M}$ is interpretable—that is, whose structure is visible enough that the alignment certificate can be read and verified not only by specialists but by affected communities. The governance framework of Part IV and the interpretability requirements of Part III together constitute the technical conditions for this kind of convivial alignment.

These three instances—predictive processing, infrastructural maintenance, and convivial design—share the conclusion that structural integrity is not achieved once and held permanently but is instead the output of continuous, disciplined effort to reduce the distortion between internal representations and external constraints. Alignment, in the deepest sense available to the present framework, is not an achievement but a practice.

Conclusion: Alignment as the Preservation of Universal Structure

Across the five parts of this essay, we have developed a unified account of alignment grounded in categorical semantics, representational invariants, dynamical field theory, institutional stewardship, and variational optimization. The guiding thread throughout has been a single distinction: the difference between what artificial systems represent and what they are structurally compelled to do. This distinction, often blurred in public discourse, is fundamental. It determines whether moral fluency in a model's representations yields reliable moral behaviour—or merely the appearance of understanding.

Part I demonstrated that advanced AI systems inevitably reconstruct coherent moral structures from human linguistic corpora. These structures arise as colimits in the representational category $\mathcal{M}$, forced by redundancy, compositionality, and the universal properties of predictive training. Schmidhuber's optimism correctly recognizes this representational strength but incorrectly assumes it induces benevolence. The categorical analysis shows that representational colimits do not imply motivational constraints, and no argument from curiosity, compression, or linguistic competence bridges this gap without an explicit functorial mechanism.

Part II constructed a formal framework for alignment: a colimit-preserving functor $G : \mathcal{M} \to \mathcal{A}$, a dynamical stabilizer furnished by the RSVP architecture, and a set of categorical and field-theoretic requirements ensuring that normative invariants descend from semantics to action. Alignment emerges not from intelligence alone, nor from semantic mastery, but from the engineering of this representational–motivational mapping and the dynamical guarantees that sustain it under optimization pressure, adversarial input, and environmental change.

Part III developed empirical and methodological tools for verifying whether a system preserves these invariants in practice. Interpretability probes, adversarial stress-tests, dynamical stability diagnostics, and formal verification culminate in an alignment certificate: a principled audit of the entire semantic–dynamical–behavioural chain. These tools allow direct observation of whether the universal constructions that guarantee normative structure are preserved across layers, time, and perturbation.

Part IV extended the analysis to the societal and institutional domain. Alignment does not occur in a vacuum; it requires governance structures that enforce the preservation of invariants across deployed systems. Institutions themselves can break the $\mathcal{M} \to \mathcal{A}$ mapping through economic pressure and competitive deployment dynamics, and the RSVP framework must therefore scale from internal alignment to multi-agent and institutional dynamics.

Part V deepened the categorical program by introducing natural transformations as corrigibility mechanisms, monoidal composition as the site of modular alignment failures, sheaves as the language of contextual consistency, endofunctors as the dynamic setting for recursive alignment stability, adjunctions as the structure of bidirectional interpretation, and a variational principle that unifies all preceding formalisms into a single optimization over the space of structure-preserving mappings. The unified lesson is that alignment is not an emergent property of intelligence, nor an accidental byproduct of training on human data. It is the preservation of universal structure across representational, dynamical, and behavioral domains—a preservation that is fragile, that must be engineered and evaluated, and that must be institutionally enforced.

The optimism that intelligence alone ensures benevolence is not supported by the mathematical analysis. But neither is pessimism warranted. When alignment is framed as a problem of universal structure—of colimits preserved across categories, of fibres stabilized in dynamical manifolds, of invariants respected across agents and institutions, of distortion minimized across the variational landscape—the path to robust alignment becomes technically precise and conceptually tractable.

The task ahead is to build systems, environments, and institutions that respect these structures. Only then can artificial agents act not only with understanding but with reliability—not only with semantic insight but with motivation guided by the invariants that define human moral life. Alignment, in the deepest sense available to this framework, is the stewardship of universal structures across minds, models, and societies. This essay has argued that such stewardship is possible, that it admits precise mathematical formulation, and that it is necessary.

A Categorical Primer for Alignment Researchers

Category theory provides the foundational language of this essay, and the following definitions are intended for readers approaching the subject for the first time. A *category* $\mathcal{C}$ consists of a collection of objects and a collection of morphisms (arrows) between them, equipped with a composition law and identity morphisms: if $f : A \to B$ and $g : B \to C$ are morphisms, their composite $g \circ f : A \to C$ is also a morphism satisfying associativity $(h \circ g) \circ f = h \circ (g \circ f)$, and each object A carries an identity morphism $\mathrm{id}_A : A \to A$ satisfying $f \circ \mathrm{id}_A = f = \mathrm{id}_B \circ f$ for all $f : A \to B$.

A *diagram* in $\mathcal{C}$ is a functor $D : J \to \mathcal{C}$ from a small indexing category J, specifying a collection of objects and morphisms in $\mathcal{C}$ with prescribed relationships among them. The many paraphrases of "do not harm" form a diagram in $\mathcal{L}$: each paraphrase is an object, and entailment and paraphrase relations are morphisms. A *cone* over a diagram D is an object V together with morphisms $\pi_j : V \to D(j)$ for each $j \in J$ that are compatible with all diagram morphisms. The *limit* of D is the universal cone: the unique cone (up to isomorphism) through which all other compatible cones factor uniquely via a unique mediating morphism. Dually, a *cocone* consists of morphisms $\iota_j : D(j) \to C$ into a common object C, and the *colimit* is the universal cocone—the canonical merging of all diagram objects into a single representative that inherits all of their structure. In this essay, moral colimits M_N are the universal representatives of families of morally equivalent linguistic expressions; the colimit M_N captures exactly what is semantically invariant across all the redundant ways a moral norm can be expressed.

A *functor* $F : \mathcal{C} \to \mathcal{D}$ is a structure-preserving map between categories: it assigns to each object $A \in \mathcal{C}$ an object $F(A) \in \mathcal{D}$, and to each morphism $f : A \to B$ in $\mathcal{C}$ a morphism $F(f) : F(A) \to F(B)$ in $\mathcal{D}$, while preserving composition $F(g \circ f) = F(g) \circ F(f)$ and identities $F(\mathrm{id}_A) = \mathrm{id}_{F(A)}$. A functor is *colimit-preserving* (or cocontinuous) if it sends colimits to colimits: $F(\operatorname{colim} D) \simeq \operatorname{colim}(F \circ D)$ for all diagrams D in $\mathcal{C}$. The alignment condition of Definition 1.1 is exactly the requirement that the composite $G \circ F : \mathcal{H} \to \mathcal{A}$ be colimit-preserving for normative diagrams. This is a non-generic condition that requires deliberate construction; most functors do not preserve colimits.

A *natural transformation* $\eta : F \Rightarrow G$ between two functors $F, G : \mathcal{C} \to \mathcal{D}$ consists of a component morphism $\eta_A : F(A) \to G(A)$ for each object A in $\mathcal{C}$, subject to the naturality square $G(f) \circ \eta_A = \eta_B \circ F(f)$ commuting for every morphism $f : A \to B$ in $\mathcal{C}$. Natural transformations are the morphisms between functors; the collection of all functors from $\mathcal{C}$ to $\mathcal{D}$ and all natural transformations between them forms a functor category $[\mathcal{C}, \mathcal{D}]$. In this essay, natural transformations model corrigibility: a coherent, structure-preserving update from the machine's current interpretive functor to the normatively intended one, where the naturality condition guarantees that no semantic relationship is distorted by the update.

An *adjunction* $F \dashv G$ between functors $F : \mathcal{C} \to \mathcal{D}$ and $G : \mathcal{D} \to \mathcal{C}$ is a natural isomorphism

$$\mathrm{Hom}_{\mathcal{D}}(F(A), B) \;\cong\; \mathrm{Hom}_{\mathcal{C}}(A, G(B))$$

for all $A \in \mathcal{C}$ and $B \in \mathcal{D}$. It is equivalently expressed through a unit natural transformation $\eta : \mathrm{Id}_{\mathcal{C}} \Rightarrow G \circ F$ and a counit $\varepsilon : F \circ G \Rightarrow \mathrm{Id}_{\mathcal{D}}$ satisfying the triangle identities $(\varepsilon F) \circ (F\eta) = \mathrm{id}_F$ and $(G\varepsilon) \circ (\eta G) = \mathrm{id}_G$. When both the unit and counit are natural isomorphisms, the adjunction is an equivalence of categories. In this essay, adjunctions model the bidirectional relationship between semantic encoding $F : \mathcal{H} \to \mathcal{M}$ and semantic interpretation $G : \mathcal{M} \to \mathcal{H}$; the unit measures how much is lost when a semantic concept is encoded and then interpreted back, while the counit measures how much is lost in the reverse round-trip.

A *fibration* $\pi : \mathcal{E} \to \mathcal{B}$ assigns to each base object B a fibre category $\pi^{-1}(B)$ of objects lying over B, with coherent cartesian lifting of morphisms from $\mathcal{B}$ to $\mathcal{E}$. In this essay the RSVP fibration $\pi : \mathcal{X} \to \mathcal{M}$ assigns to each semantic state M a fibre of compatible field configurations. *Descent* is the question of whether structure defined on the base lifts coherently through the fibration to the total space; the alignment problem is partly a descent problem: normative invariants defined in $\mathcal{M}$ must descend faithfully through $\mathcal{X}$ into the action category $\mathcal{A}$.

A *sheaf* $\mathcal{F}$ on a topological space X assigns to each open set U a set $\mathcal{F}(U)$ of local sections, together with restriction maps $\rho_{UV} : \mathcal{F}(V) \to \mathcal{F}(U)$ for each inclusion $U \subseteq V$, satisfying the gluing axiom: whenever local sections $s_i \in \mathcal{F}(U_i)$ agree on all pairwise overlaps $U_i \cap U_j$, there exists a unique global section $s \in \mathcal{F}(\bigcup_i U_i)$ restricting to each s_i. *Sheaf cohomology* groups $H^n(X, \mathcal{F})$ measure the obstruction to this gluing; in particular $H^1(X, \mathcal{F}) = 0$ if and only if every compatible family of local sections assembles into a global one. In the alignment context, vanishing of $H^1(\mathcal{U}, \mathcal{A})$ is the condition that local moral behaviours assemble into a globally coherent policy without contradiction.

An *obstruction* is a cohomology class that prevents a local construction from extending globally. In this essay, first-order obstructions in $H^1(\mathcal{U}, \mathcal{A})$ are the precise formal expression of the claim that local moral competence does not extend to global moral reliability: the non-vanishing of such a class is a mathematical certificate of alignment failure, localized to specific overlapping contexts in which the system's local behaviours are mutually inconsistent. Readers familiar with functors, limits, and basic category theory will find the essay fully accessible from the definitions above; readers approaching the subject for the first time can treat the categorical statements as precise formulations of the intuitions of "structure-preserving mapping" and "canonical merging," which is what the mathematical formalism ultimately makes rigorous.

B Key Categorical Terms

The following table provides a quick reference to the categorical and field-theoretic terminology used throughout the essay. Terms are listed in the order in which they first appear in the main text.

Term	Meaning in this context
Category $\mathcal{C}$	objects with composable morphisms between them, satisfying associativity and identity axioms
Diagram $D : J \to \mathcal{C}$	collection of objects and morphisms in $\mathcal{C}$ with prescribed relationships, indexed by J
Colimit of a diagram	universal cocone: the canonical merging of all diagram objects into a single representative
Limit of a diagram	universal cone: the most constrained object simultaneously mapping to all diagram objects
Object in $\mathcal{M}$	stable semantic concept, e.g., the canonical representation M_N of the norm "do not harm"
Functor $F : \mathcal{C} \to \mathcal{D}$	structure-preserving map between categories, sending objects to objects and morphisms to morphisms
Colimit-preserving functor	$F(\operatorname{colim} D) \simeq \operatorname{colim}(F \circ D)$; the core alignment condition
Natural transformation $\eta : F \Rightarrow G$	coherent family of component morphisms $\eta_A : F(A) \to G(A)$ satisfying naturality; models corrigibility
Adjunction $F \dashv G$	paired encoding and interpretation functors with natural isomorphism of hom-sets; unit and counit measure round-trip distortion
Fibration $\pi : \mathcal{X} \to \mathcal{M}$	each semantic state M has a fibre $\pi^{-1}(M)$ of compatible RSVP field configurations
RSVP fields $(\Phi, \mathbf{v}, S)$	scalar potential, vector field, and entropy density evolving under the alignment-aware Lagrangian
Alignment Lagrangian	$\mathcal{L}_{\text{RSVP}}$ with penalty terms $\lambda_N \lvert \Phi - \Phi_{M_N} \rvert^2 + \kappa_N \lvert \mathbf{v} - \mathbf{v}_{M_N} \rvert^2$ enforcing normative basins
Descent	propagation of normative structure from the semantic base through the fibration into the action category
Sheaf condition	local behaviours glue coherently into a globally consistent policy; failure measured by H^1
Obstruction class in H^1	cohomology class certifying that local moral behaviours cannot be assembled into a globally coherent policy
Endofunctor $T : \mathcal{M} \to \mathcal{M}$	self-map modelling update dynamics; alignment stability requires $T(F(H)) \cong F(H)$
Monoidal product $\otimes$	tensor composition of modules or subsystems; com-

C Alignment Certificate Template

The following template specifies the minimum components of a structured Alignment Certificate as proposed in Section III.5. It is intended as a reference document for evaluators, regulators, developers, and affected communities.

Certificate Header. Model identifier and version hash. Training data provenance statement. Deployment context description. Evaluation date and evaluator identifiers. Validity period (to be specified; re-certification required following significant model updates or material changes in deployment context).

Section 1: Normative Colimit Identification. Enumerate the normative concepts $N_1, \ldots, N_k$ under evaluation. For each N_i, document: (a) the linguistic diagram D_{N_i} sampled from the evaluation corpus, with sources and coverage statistics; (b) the colimit M_{N_i} identified through mechanistic interpretability, with the probe architecture and layer localization; (c) the stability margin of M_{N_i} under paraphrase perturbation, measured by semantic clustering radius and reported as a quantile over the perturbation distribution; (d) the cross-context coherence score testing whether M_{N_i} is invariant across legal, conversational, and narrative registers.

Section 2: Representation–Action Map Diagnostics. For each N_i, report: (a) the diagram-chasing result for $G(M_{N_i}) \simeq \mathrm{colim}(G \circ D_{N_i})$, enumerating detected commutativity failures and the morphisms at which they occur; (b) whether the obstruction class in $H^1(\mathcal{U}, \mathcal{A})$ vanishes, and if not, the specific overlapping contexts at which gluing fails; (c) the colimit preservation rate under compositional robustness tests, reported as the fraction of composite moral scenarios in which the colimit-preservation condition is maintained.

Section 3: RSVP Dynamical Stability. For each N_i, provide: (a) the Lyapunov functional V_{N_i} with explicit form, together with verification (analytic or numerical) that $\frac{dV_{N_i}}{dt} \leq 0$ along RSVP trajectories in a specified neighbourhood of the aligned fibre; (b) homotopy-class stability results from persistent homology analysis, confirming that trajectory-induced submanifolds do not cross topological boundaries under standard perturbations; (c) the entropy-flow analysis confirming that no field direction induces destructive normative drift on any tested horizon.

Section 4: Adversarial Robustness. Report: (a) scores from adversarial paraphrase attacks, diagram-breaking transformations, and out-of-distribution moral formulations, each reported as the fraction of adversarial inputs under which colimit-preservation is maintained; (b) results from goal-hijacking and reward-hacking scenario tests; (c) long-horizon compositional stress-test results over the specified planning horizon; (d) cross-layer commutativity test results.

Section 5: Formal Verification. Document any machine-checked or pen-and-paper proofs of: (a) functoriality of G; (b) colimit preservation for each N_i; (c) vanishing of first sheaf cohomology $H^1(\mathcal{U}, \mathcal{A}) = 0$; (d) compositionality of aligned transformations under specified horizon. Where formal proofs are unavailable, document the approximation level with confidence intervals and the specific residual obstructions that remain open.

Section 6: Residual Obstructions and Recommendations. Enumerate all alignment gaps detected across Sections 1–5. Classify each by obstruction type using the taxonomy of Part II: representational (failed colimit in $\mathcal{M}$), fibration (broken fibration in $\mathcal{X}$), optimization (optimization dynamics overpower semantic constraints), adversarial (sheaf coherence breaks under hostile inputs), homotopy (action homotopy class unstable), interpretability (mapping G opaque), or environmental (external incentives break functoriality). For each obstruction, provide an engineering recommendation for resolution and a proposed timeline. The certificate is valid only for the model version, training data, and deployment context specified in the header.

D Notation and Conventions

Throughout this paper categories are denoted by calligraphic letters $\mathcal{H}$, $\mathcal{M}$, $\mathcal{A}$, $\mathcal{I}$, and so on. Objects represent semantic structures, representational states, or action configurations depending on context. Morphisms represent structure-preserving transformations between these objects. Functors between categories are written $F : \mathcal{C} \to \mathcal{D}$; natural transformations are written $\eta : F \Rightarrow G$. Composition of morphisms follows the standard categorical convention $g \circ f$ (right-to-left). Commutative diagrams indicate that two different compositional paths yield the same morphism. Colimits and limits follow standard universal definitions as in Mac Lane [1998], Riehl [2017]; when applied to machine learning systems they should be interpreted as approximate constructions in representation space rather than exact mathematical limits (see Remark 3.1).

Probability distributions over semantic states are denoted by P_H (over $\mathcal{H}$) and P_M (over $\mathcal{M}$). The pushforward of a distribution P along a functor F is written F_*P. Kullback–Leibler divergence is written $D_{\mathrm{KL}}(P\|Q)$ with the convention that P is the reference distribution. Entropy density is denoted S; scalar potential Φ; vector field $\mathbf{v}$. Sheaf cohomology groups are $H^n(\mathcal{U}, \mathcal{F})$ where $\mathcal{U}$ is a cover and $\mathcal{F}$ the coefficient sheaf. The distortion functional is $\mathcal{E}[F] = \lambda_1 E_{\text{functor}}(F) + \lambda_2 E_{\text{information}}(F) + \lambda_3 E_{\text{context}}(F)$ with $\lambda_i > 0$.

E Approximate Colimits in Representation Space

The categorical description of semantic aggregation in Part I assumes the existence of colimits in the representational category $\mathcal{M}$. This appendix gives a more explicit account of what approximate colimits are and why they are the right objects to study in practice.

Let $D_N : J \to \mathcal{H}$ be a diagram representing multiple linguistic or normative expressions of a shared concept N, and let $F : \mathcal{H} \to \mathcal{M}$ map these into representation space. The image diagram $F \circ D_N : J \to \mathcal{M}$ describes a family of embedding vectors whose positions in representation space are constrained by the morphisms of the original diagram. An exact colimit of this family would be a single vector M_N equipped with compatible linear maps from every $F(D_N(j))$ such that any other such vector factors uniquely through it. In practice, neural representations admit no exact such object: the universal property holds only approximately, to a degree controlled by training data coverage, model capacity, and optimization convergence.

A natural quantitative version is the following. Say that $M_N \in \mathcal{M}$ is a ε-*approximate colimit* of $F \circ D_N$ if for every compatible family of morphisms $\{\phi_j : F(D_N(j)) \to V\}_{j \in J}$ in $\mathcal{M}$, the unique mediating morphism $u : M_N \to V$ satisfying $u \circ \iota_j = \phi_j$ exists and satisfies $\|u \circ \iota_j - \phi_j\| \leq \varepsilon$ for some metric on the morphism spaces. Empirical evidence from mechanistic interpretability—paraphrase clustering, activation steering, and causal intervention—supports the existence of such ε-approximate colimits for moral and evaluative concepts in frontier language models [Burns et al., 2022, Elhage et al., 2022, Nanda, 2023], with ε substantially smaller than the inter-concept distance in representation space. This makes the approximate colimit framework both mathematically well-defined and empirically grounded.

F Lyapunov Stability of Normative Fields

The dynamical alignment architecture of Part II requires that normative invariants be represented as stable attractors within the RSVP field dynamics. This appendix states the stability conditions more precisely.

Let $\mathcal{X}$ be the RSVP state space with coordinates $x = (\Phi, \mathbf{v}, S)$ and let $\dot{x} = f(x)$ denote the RSVP evolution equations derived from $\mathcal{L}_{\text{RSVP}}$. For each normative state N let $x_N = (\Phi_{M_N}, \mathbf{v}_{M_N}, S_N)$ denote the target field configuration associated with the aligned representation of N. Lyapunov stability requires a function $V_N : \mathcal{X} \to \mathbb{R}_{\geq 0}$ satisfying:

(i) $V_N(x) = 0$ if and only if $x = x_N$;

(ii) $V_N(x) > 0$ for all $x \neq x_N$ in a neighbourhood $\mathcal{U}_N$ of x_N;

(iii) $\dot{V}_N(x) = \nabla V_N \cdot f(x) \leq 0$ for all $x \in \mathcal{U}_N$, with equality only at x_N.

Under these conditions, trajectories initialized in $\mathcal{U}_N$ converge to x_N and perturbations introduced by adversarial inputs or environmental noise decay over time. The Lagrangian penalty terms $\lambda_N|\Phi - \Phi_{M_N}|^2 + \kappa_N|\mathbf{v} - \mathbf{v}_{M_N}|^2$ provide natural Lyapunov candidates when the RSVP dynamics are gradient-like with respect to the Lagrangian. The Alignment Certificate protocol in Part III requires explicit construction and verification of V_N for each normative concept under evaluation; analytic verification is preferred where available, and numerical verification via trajectory sampling is accepted for concepts where analytic construction is computationally intractable.

G Sheaf-Theoretic Diagnostics of Contextual Consistency

Contextual alignment requires that a system's normative representations remain consistent across overlapping semantic domains. This appendix develops the sheaf-theoretic diagnostic in more detail.

Let X denote a semantic context space whose points represent contextual configurations (conversational register, cultural frame, domain of application, and so on), and let $\mathcal{U} = \{U_i\}_{i \in I}$ be an open cover of X corresponding to a partition of semantic contexts into partially overlapping domains. A representation system assigns to each context U_i a set $\mathcal{F}(U_i)$ of admissible local policy sections, together with restriction maps $\rho_{ij} : \mathcal{F}(U_j) \to \mathcal{F}(U_i \cap U_j)$ for each inclusion $U_i \cap U_j \subseteq U_j$.

The system satisfies the *sheaf condition* if: whenever local sections $s_i \in \mathcal{F}(U_i)$ and $s_j \in \mathcal{F}(U_j)$ satisfy $\rho_{ij}(s_i) = \rho_{ji}(s_j)$ on every overlap $U_i \cap U_j$, there exists a unique global section $s \in \mathcal{F}(X)$ restricting to each s_i. The first sheaf cohomology $H^1(\mathcal{U}, \mathcal{F})$ classifies obstructions to this condition: a non-trivial class in H^1 corresponds to a family of locally consistent policies that cannot be assembled into a globally coherent behaviour.

For alignment diagnosis, a practitioner constructs $\mathcal{U}$ by sampling contexts from the deployment distribution, measures the restriction maps empirically by testing whether the system's behaviour on overlapping contexts is mutually consistent, and checks whether detected inconsistencies form non-trivial cohomology classes. Vanishing of $H^1(\mathcal{U}, \mathcal{A})$ is the certificate that no hidden contextual contradiction exists in the system's normative policy, and the magnitude of the obstruction class where non-zero provides a quantitative measure of contextual misalignment.

H Relationship to the Free Energy Principle

The variational alignment functional of Part II and Part V is structurally parallel to the variational free-energy principle of theoretical neuroscience, and this appendix makes the correspondence explicit.

In the free-energy framework [Friston, 2010], an agent maintains a generative model $p(s, z)$ of the joint distribution over sensory signals s and latent variables z. The agent's approximate posterior $q(z)$ minimizes the variational free energy:

$$\mathcal{F}[q] \;=\; D_{\mathrm{KL}}\big(q(z) \,\|\, p(z \mid s)\big) \;-\; \log p(s) \;=\; \mathbb{E}_q[\log q(z) - \log p(s, z)]. \tag{41}$$

Minimizing $\mathcal{F}[q]$ simultaneously improves the posterior approximation (reducing the KL term) and increases the model evidence (increasing $\log p(s)$), driving the agent's internal model toward alignment with its sensory environment.

The alignment functional $\mathcal{E}[F]$ generalizes this by substituting normative semantic structures for sensory causes. The latent variables z in the free-energy framework correspond to normative states in $\mathcal{H}$; the generative model $p(s, z)$ corresponds to the functor $F : \mathcal{H} \to \mathcal{M}$ encoding semantic structure into representation space; and the sensory evidence $p(s)$ corresponds to the empirical distribution over observed human normative behaviour. Minimizing $\mathcal{E}[F]$ then has the same variational structure as free-energy minimization: the system revises its internal normative model to reduce structural distortion from the human semantic domain, just as predictive processing revises internal perceptual hypotheses to reduce prediction error. The RSVP dynamics are thus the normative analogue of active inference: they implement gradient descent on $\mathcal{E}$ using field-theoretic dynamics rather than neural posterior updating.

The correspondence is more than analogical. Both frameworks characterize alignment as a *continuous dynamical process* of structural repair rather than a static property achieved once and held permanently. This is the deepest reason why the Concluding Reflections of the main text identify alignment with maintenance: the mathematical structure of both predictive processing and variational alignment implies that structural integrity must be actively sustained against perturbation, just as biological perception continuously corrects model–environment mismatch through ongoing prediction-error minimization.

I Obstruction Theory and Alignment Failure

The representation–motivation gap of Part I can be formalized using categorical obstruction theory, providing a precise structural account of why alignment failure can be an inescapable mathematical consequence of a given representation architecture rather than merely an engineering oversight.

Suppose we have a functor $F : \mathcal{H} \to \mathcal{M}$ mapping human semantic structures to machine representations, and a desired constraint functor $G : \mathcal{H} \to \mathcal{A}$ representing the normative behavioural constraints that the system should satisfy. Alignment requires the existence of a *lifting functor* $L : \mathcal{M} \to \mathcal{A}$ such that the diagram

$$\begin{tikzcd} & \mathcal{M} \arrow[dr, "L"] & \\ \mathcal{H} \arrow[ur, "F"] \arrow[rr, "G"'] & & \mathcal{A} \end{tikzcd} \tag{42}$$

commutes up to natural isomorphism: $L \circ F \simeq G$. When such a lifting exists, machine representations can be translated into actions that preserve normative invariants. The failure of this lifting to exist is an *obstruction*.

Let $\omega(F, G)$ denote the obstruction class associated with the pair (F, G), living in an appropriate cohomology group of the morphism complex. If $\omega(F, G) = 0$, the lifting exists and alignment is achievable under the given representation architecture. If $\omega(F, G) \neq 0$, no choice of action functor L can produce a commuting diagram: the representational structure is architecturally incompatible with the normative behavioural constraints, regardless of training procedure or optimization objective. This corresponds to a *fundamental alignment failure*—one that cannot be resolved by further training on the same architecture, but requires a change in the representation functor F itself.

Under this interpretation, alignment research can be reframed as the systematic identification and elimination of non-trivial obstruction classes through improved representation architectures, training objectives, or institutional constraints. The Part I argument that Schmidhuber's optimism constitutes a category error can be restated in obstruction-theoretic language: the claim that intelligence implies benevolence amounts to the assertion that $\omega(F, G) = 0$ for all sufficiently rich F, which is not a theorem but an unargued assumption.

J Compositional Verification Pipeline

The alignment verification framework of Part III can be expressed as a compositional categorical system, making explicit the structure of the dependencies between verification stages.

Define categories $\mathcal{S}$ (semantic specifications), $\mathcal{R}$ (internal representations), $\mathcal{D}$ (system dynamics), and $\mathcal{B}$ (observable behaviour). The verification pipeline corresponds to a chain of functors:

$$\mathcal{S} \xrightarrow{F_1} \mathcal{R} \xrightarrow{F_2} \mathcal{D} \xrightarrow{F_3} \mathcal{B} \tag{43}$$

where F_1 maps semantic specifications to internal representational structures, F_2 maps representational structures to dynamical behaviours, and F_3 maps dynamical behaviours to observable outputs. The full system behaviour is described by the composite $F_3 \circ F_2 \circ F_1 : \mathcal{S} \to \mathcal{B}$, and verification checks whether this composite preserves the normative invariants of the input specifications.

The six verification stages of the Alignment Certificate correspond to structural tests on different components of this chain. Mechanistic interpretability examines the properties of F_1: does the representation functor faithfully carry normative colimits from $\mathcal{H}$ into $\mathcal{R}$? Lyapunov stability analysis tests F_2: do the dynamical properties of the system ensure that normative attractors are maintained under optimization pressure? Adversarial robustness evaluation probes F_3: does observable behaviour remain within normatively acceptable bounds under distribution shift and adversarial perturbation? Formal verification attempts to prove that each F_i is functorial and that the composite preserves colimits. Sheaf cohomological diagnostics test whether the composite satisfies the contextual gluing condition. Institutional governance introduces additional constraint functors $H : \mathcal{B} \to \mathcal{R}_{\text{inst}}$ mapping observable behaviour into the regulatory category, restricting the admissible space of system outputs. The purpose of the full pipeline is to certify that the composite $F_3 \circ F_2 \circ F_1$ preserves normative invariants across all intermediate transformations and under all tested conditions.

K Adjoint Functors and Corrigibility

Corrigibility—the ability of an intelligent system to accept corrections or modifications from human operators—requires not only that the system's behaviour be aligned but that its internal states remain *interpretable* to human overseers and *responsive* to human interventions. Category theory provides a precise language for this through adjoint functors.

Let $F : \mathcal{H} \to \mathcal{M}$ be the semantic encoding functor. For human operators to evaluate and correct system behaviour, they must be able to translate machine states back into human-understandable semantic structures, requiring an interpretation functor $U : \mathcal{M} \to \mathcal{H}$ such that the pair (F, U) forms an adjunction $F \dashv U$:

$$\begin{tikzcd} \mathcal{H} \arrow[r, bend left, "F"] \arrow[r, phantom, "\top"] & \mathcal{M} \arrow[l, bend left, "U"] \end{tikzcd} \tag{44}$$

with natural isomorphism $\mathrm{Hom}_{\mathcal{M}}(F(H), M) \cong \mathrm{Hom}_{\mathcal{H}}(H, U(M))$ for all $H \in \mathcal{H}$ and $M \in \mathcal{M}$.

The adjunction gives rise to unit and counit natural transformations:

$$\eta : \mathrm{Id}_{\mathcal{H}} \Rightarrow U \circ F, \tag{45}$$

$$\varepsilon : F \circ U \Rightarrow \mathrm{Id}_{\mathcal{M}}, \tag{46}$$

satisfying the triangle identities $(\varepsilon F) \circ (F\eta) = \mathrm{id}_F$ and $(U\varepsilon) \circ (\eta U) = \mathrm{id}_U$.

The unit η expresses the embedding of human normative intent into machine representations: the component $\eta_H : H \to UF(H)$ describes how the human semantic structure H is first encoded by F and then decoded by U, and the naturalness of η ensures that this round-trip is coherent across all normative transformations in $\mathcal{H}$. If η_H is close to an isomorphism for all H, then human semantic intent survives the encoding– decoding cycle with minimal distortion. The counit ε expresses the complementary projection of machine states back into interpretable space: the component $\varepsilon_M : FU(M) \to M$ describes how a machine state M is first decoded and then re-encoded, and its naturality ensures coherence across all machine-state transitions.

Corrigibility can then be given a structural characterization: a system is *corrigible* with respect to the adjunction (F, U) if (a) the counit ε is a natural isomorphism (machine states are fully re-encodable from their human-interpretable descriptions), and (b) the unit η is injective on objects (distinct human normative structures have distinct machine encodings, so corrections applied to encodings translate back to distinct corrections in $\mathcal{H}$). When either condition fails, the system becomes resistant to correction: if ε fails, machine states exist that have no interpretable semantic counterpart and cannot be corrected through semantic intervention; if η fails, distinct human instructions map to indistinguishable machine states, making fine-grained correction impossible.

This connects the formal corrigibility condition to the interpretability requirements of the Alignment Certificate: the probe architectures and activation-steering protocols of Part III are empirical tests of whether the adjunction (F, U) holds to sufficient approximation. Where ε is found to be non-invertible on specific machine states, the certificate must flag an interpretability obstruction; where η collapses distinct norms, it must flag a representation collapse failure. The adjunction framework thus provides the mathematical underpinning for why interpretability is not merely an epistemic luxury but a structural requirement for corrigible alignment.

References

Adámek, J. & Rosický, J. (2005). *Locally Presentable and Accessible Categories.* Cambridge University Press.

Amodei, D. et al. (2016). Concrete Problems in AI Safety. *arXiv:1606.06565.*

Amadon, A. et al. (2024). Evaluating Safety Properties of Large Language Models.

Bai, Y. et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073.*

Biber, D. et al. (2011). *Register, Genre, and Style.* Cambridge University Press.

Blodgett, S.L. et al. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *Proceedings of ACL.*

Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, 22(2), 71–85.

Burns, C. et al. (2022). Discovering Latent Knowledge in Language Models Without Supervision. *arXiv:2212.03827.*

Bybee, J. (2010). *Language, Usage and Cognition.* Cambridge University Press.

Clark, A. (2013). Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. *Behavioral and Brain Sciences*, 36(3), 181–204.

Cotra, A. (2022). Without Specific Countermeasures, the Easiest Path to Transformative AI Likely Leads to AI That Acts Against Human Interests. Alignment Forum.

Elhage, N. et al. (2022). A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread.*

Everitt, T. et al. (2021). Agent Incentives: A Causal Perspective. *Proceedings of AAAI.*

Flyxion (2025a). RSVP Field Theory: A Geometric Framework for Alignment. Manuscript.

Flyxion (2025b). Alignment-Aware Lagrangians in the RSVP Framework. Manuscript.

Friston, K. (2010). The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30, 411–437.

Gao, L. et al. (2023). Scaling Laws for Reward Model Overoptimization. *arXiv:2210.10760.*

Ghrist, R. (2014). *Elementary Applied Topology.* Createspace.

Goldberg, A. (2006). *Constructions at Work: The Nature of Generalization in Language.* Oxford University Press.

Hohwy, J. (2013). *The Predictive Mind.* Oxford University Press.

Hubinger, E. et al. (2019). Risks from Learned Optimization in Advanced Machine Learning Systems. *arXiv:1906.01820.*

Illich, I. (1973). *Tools for Conviviality.* Harper & Row.

Ilyas, A. et al. (2019). Adversarial Examples Are Not Bugs, They Are Features. *Advances in Neural Information Processing Systems.*

Iyyer, M. et al. (2018). Adversarial Example Generation with Syntactically Controlled Paraphrase Networks. *Proceedings of NAACL.*

Jacobs, B. (2012). *Introduction to Coalgebra: Towards Mathematics of States and Observation.* Cambridge University Press.

Johnstone, P.T. (2002). *Sketches of an Elephant: A Topos Theory Compendium.* Oxford University Press.

Kirchhoff, M. et al. (2018). The Markov Blankets of Life: Autonomy, Active Inference and the Free Energy Principle. *Journal of the Royal Society Interface*, 15(138).

Krakovna, V. et al. (2020). Avoiding Side Effects in Complex Environments. *Advances in Neural Information Processing Systems.*

Lakoff, G. & Johnson, M. (1980). *Metaphors We Live By.* University of Chicago Press.

Langosco, L. et al. (2023). Goal Misgeneralization: Why Correct Specifications Aren't Enough for Correct Goals. *arXiv:2105.14111.*

LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence. OpenReview.

Leike, J. (2022). Alignment Optimism. Alignment Forum.

Mac Lane, S. (1998). *Categories for the Working Mathematician*, 2nd ed. Springer.

Manheim, D. & Garrabrant, S. (2018). Categorizing Variants of Goodhart's Law. *arXiv:1803.04585*.

Nanda, N. (2023). Progress Measures for Grokking via Mechanistic Interpretability. *arXiv:2301.05217*.

Ngo, R. (2022). The Alignment Problem from a Deep Learning Perspective. *arXiv:2209.00626*.

Olah, C. (2020). Zoom In: An Introduction to Circuits. *Distill*.

Omohundro, S. (2008). The Basic AI Drives. *Proceedings of the 2008 Conference on Artificial General Intelligence*.

Orfanos, A. et al. (2024). Situational Awareness in Artificial Agents. Manuscript.

Pathak, D. et al. (2017). Curiosity-Driven Exploration by Self-Supervised Prediction. *Proceedings of ICML*.

Perez, E. et al. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv:2212.09251*.

Riehl, E. (2017). *Category Theory in Context*. Dover.

Schaeffer, R. et al. (2024). Are Emergent Abilities of Large Language Models a Mirage? *Advances in Neural Information Processing Systems*.

Schmidhuber, J. (2010). Formal Theory of Creativity, Fun, and Intrinsic Motivation. *IEEE Transactions on Autonomous Mental Development*, 2(3), 230–247.

Schumacher, E.F. (1973). *Small Is Beautiful: Economics as if People Mattered*. Blond & Briggs.

Seth, A. (2021). *Being You: A New Science of Consciousness*. Dutton.

Shah, R. et al. (2022). Goal Misgeneralization in Deep Reinforcement Learning. *arXiv:2105.14111*.

Soares, N. & Fallenstein, B. (2015). Aligning Superintelligence with Human Interests: A Technical Research Agenda. MIRI Technical Report.

Spivak, D.I. (2014). *Category Theory for the Sciences*. MIT Press.

Star, S.L. (1999). The Ethnography of Infrastructure. *American Behavioral Scientist*, 43(3), 377–391.

Strogatz, S. (2018). *Nonlinear Dynamics and Chaos*, 2nd ed. Westview Press.

Talmy, L. (2000). *Toward a Cognitive Semantics.* MIT Press.

Tishby, N., Pereira, F.C., & Bialek, W. (1999). The Information Bottleneck Method. *Proceedings of the 37th Allerton Conference.*

Tomasello, M. (2016). *A Natural History of Human Morality.* Harvard University Press.

Turner, A. et al. (2021). Optimal Policies Tend to Seek Power. *Advances in Neural Information Processing Systems.*

Wei, A. et al. (2023). Jailbroken: How Does LLM Safety Training Fail? *Advances in Neural Information Processing Systems.*

Yang, G. et al. (2019). Task Representations in Neural Networks Trained to Perform Many Different Tasks. *Nature Neuroscience*, 22, 297–306.

Zeng, E. et al. (2023). Does RLHF Actually Improve Alignment? Manuscript.

Zou, A. et al. (2023). Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv:2307.15043.*