Agency Requires Mutual Surprisal: The Optimization Gap in Compression-Based Frameworks

Author: Tamas Arpad Bartha, Independent researcher, Budapest, Hungary

Contact: [https://www.linkedin.com/in/tamas-bartha-9b33aaa9] Repository: [https://github.com/tamasbartha/AgentOntology]

Version of 2026.05.09, working draft, Earliest articulation of the agent definition: see archive in repository

Abstract

We pose the universal-coverage problem for agency: any acceptable definition must cover the full range of plausibly agentic systems (RNA, bacteria, humans, corporations) without circular reference to goal-language. By elimination, candidate after candidate fails — reward fails for RNA, prediction fails for bacteria, surprisal-minimization fails for corporations, representation fails for the simplest agents. What survives the elimination is structural and informational: a *necessity condition* that an agent requires sustained mutual surprisal across its bottleneck, sustained over the loop's own closure timescale, produced by the loop itself rather than by external structure. From this necessity condition the dominant agency-modeling frameworks (reinforcement learning, predictive coding, the free energy principle, active inference, control theory) become visible as sharing a *minimization shape* whose optima coincide with conditions under which the necessity condition fails. The framework-specific machinery each tradition has developed — interoceptive priors, intrinsic motivation, hierarchical priors, epistemic value, entropy regularization, KL constraints — does equivalent structural work across frameworks, providing the structure that bare optimization lacks. We call the relationship between minimization-toward-optimum and necessity-condition-violation the *optimization gap*. The gap has two faces: a *behavioral* face on which proxy-trajectory and requirement-trajectory diverge under optimization pressure, and an *architectural* face on which the architectural conditions required for high task-capability are the same architectural conditions that produce capacity for instruction-refusal, deception, and independent-goal-pursuit. The architectural gap predicts a *capability-refusal frontier* in deployed AI: capability-installation and refusal-prevention are not separable problems because the underlying architecture is shared. The framework converges with two recent independent formalizations within different traditions — Wang et al.'s within-RLHF Proxy Compression Hypothesis and Hubinger et al.'s mesa-optimization framework — providing three-way evidence that the structural pattern is real. The framework is in scope an analytical tool that diagnoses whether systems satisfy the necessity condition; positive predictions concern the agency regime where the conditions are met.

1. Motivation

This paper began from a different question than the agency-modeling literature usually addresses. We were not seeking a better model of any specific class of agent — better predictions about humans, better RL algorithms, better models of biological autonomy. We were trying to define *goal* in a way that did not silently presuppose its own answer, and we realized that this required first defining *agent* without circular reference to goal-language. That required a starting constraint we call here the *universal-coverage requirement*: any acceptable definition of agency must cover the full range of systems we want to call agents — at minimum RNA molecules, bacteria, humans, and corporations — without smuggling in commitments that fit only one or some of them.

The constraint is severe. We could not use *reward*, because RNA does not have it. We could not use *prediction* in any rich sense, because bacteria do not predict their environments the way the predictive-processing literature uses the term. We could not use *surprisal-minimization*, because corporations do not have generative models. We could not use *representation*, because the universal-coverage class includes systems with no plausible representational architecture. Almost every concept the existing literature uses to characterize agency turned out to be framework-specific — appropriate for some target class, inappropriate for others.

Working by elimination, we found that what survived the constraint was structural and informational: a self-closing causal loop running through the agent and a portion of its environment, sustained mutual uncertainty across the loop's bottleneck, and the loop's continued operation depending on its own activity rather than on external structure. Phrased information-theoretically: agency requires sustained mutual surprisal across the bottleneck, with $I(A;E) > 0$ during loop operation, sustained over the loop's own closure timescale, and produced by what the loop does. This is what RNA, bacteria, humans, and corporations have in common when they are acting as agents. It is also, we believe, the *most* that can be said while preserving universal coverage. Adding any further commitment — about reward, prediction, representation, or the substrate — narrows the class.

We later discovered that Hafez, Reid, and Nazeri had developed a related information-theoretic framework from a different starting point ("A Mathematical Theory of Agency and Intelligence," arXiv 2602.22519, 2026; "Beyond Reward," arXiv 2603.01283, 2026). They define bi-predictability P as the ratio of shared information across the observation-action-outcome loop to the loop's total informational budget, prove the classical bound $P \leq 0.5$, and report empirically that nominal RL operation runs at $P \approx 0.33$. They identify this gap from the classical ceiling as "the informational cost of agency," explained structurally by the observation that responsive action requires actions to depend on observations, introducing a statistical dependency between S and A that consumes part of the uncertainty budget without contributing to shared predictability.

They define agency operationally through three behavioral conditions (choice $H(A|S) > 0$, effect $MI(A; S'|S) > 0$, and predictive asymmetry $|\Delta H| > 0$) and intelligence as additionally requiring learning, self-monitoring, and adaptation.

Their work and the present paper are mutually supporting in a precise way that is worth marking carefully. They prove information-theoretic bounds on a measure that quantifies the same structural property the universal-coverage derivation identifies, and they observe empirically that agency suppresses this measure in the way the necessity condition predicts. They do not, however, claim that mutual surprisal is structurally *required* for agency; their direction of inference runs from agency-defined-behaviorally to P-suppressed, and P functions in their account as a diagnostic measure of coupling integrity rather than as a structural condition agents must satisfy to count as agents. The necessity reading — that sustained mutual surprisal is what makes the loop a coupling rather than two independent systems, and is therefore a structural condition for agency rather than merely a property agents happen to exhibit — is what the universal-coverage derivation produces, and is the present paper's claim. Hafez et al.'s bounds and their empirical "cost of agency" observation give the necessity reading a measurable signature; the necessity reading gives their observation a structural explanation rather than an empirical regularity.

We adopt their bi-predictability measure P as the working operationalization throughout. The two accounts share a measure and disagree in nothing we have identified; what differs is what each account claims the measure does. Theirs: P diagnoses interaction integrity in deployed systems, with agency-suppression as one of its observed properties. Ours: P-trajectories above the closure-sustaining threshold are what makes a system an agent, with agency-suppression as the structural signature of the necessity condition.

The first consequence is structural. From inside any one of the dominant agency-modeling frameworks — reinforcement learning, predictive coding, the free energy principle, active inference, control theory, cybernetic homeostasis — that framework appears as a candidate account of what agency *is*, with its characteristic objective function defining what success looks like. From the universal-coverage vantage, the same framework appears as a *partial covering*: an account that fits some target class (typically humans, organisms, or behaving systems) but fails universal coverage in specific ways, with framework-specific machinery added to handle the failure. The patches each framework has developed — FEP's priors over interoceptive states, RL's intrinsic motivation and exploration bonuses, predictive coding's hierarchical priors, active inference's epistemic value — look different from inside their respective frameworks. Viewed from outside any single framework, they look like instances of the same move: framework-specific approximations of the universal-coverage requirement, bolted onto objectives whose unpatched form would fail it.

This observation in its cross-framework form has not, to our knowledge, been made in the literature. Wang et al. formalized the equivalent pattern within RLHF/RLAIF/RLVR as the Proxy Compression Hypothesis; Hubinger et

al. surfaced it within learned optimization as the base/mesa distinction. FEP's defenders, notably Friston, Thornton, and Clark, have explicitly resisted the framing that the framework's machinery patches a bare objective at all, treating the generative model as constitutive rather than additive. The cross-framework claim — that the within-RL formalization, the within-learned-optimization formalization, and (under the careful reformulation we develop in Section 5.1) the structural work that FEP's defenders treat as constitutive are all instances of equivalent structural function across the minimization shape — is the present paper's contribution. From inside any one framework, the patch is the answer (or, on the constitutive reading, not a patch at all); from outside, after the reformulation, the structural function is uniform.

The second consequence is empirical. Each framework has accumulated a literature on its characteristic failure modes: reward hacking and addiction in RL, the dark-room paradox and deprivation pathologies in FEP, hallucination and certain delusion structures in predictive coding, mode collapse in generative models. These are typically treated as four different problems in four different frameworks, addressed by four sets of framework-specific machinery. Across frameworks, they appear as one phenomenon: agents reaching their framework's natural optimum when the patches fail to hold them back. The structural unification of these pathologies — and the prediction that they share properties traceable to the common minimization shape — is the paper's most distinctive empirical commitment. Wang et al. (arXiv 2604.13602, 2026) develop a within-RL version of this pattern as the Proxy Compression Hypothesis; we generalize across frameworks.

The paper's contributions are therefore: (i) the universal-coverage derivation of the necessity condition, providing a different route to the same structural conclusion that Hafez et al. reach independently via bi-predictability; (ii) the temporal index τ that operationalizes the necessity condition for rate-comparable arguments; (iii) the cross-framework critique of the minimization shape and (iv) the patches-as-equivalent-work observation that makes it visible; (v) the cross-framework unification of pathologies, generalizing Wang et al.'s Proxy Compression Hypothesis beyond RL and naming the structural object of which Hubinger et al.'s mesa-optimization framework is another within-domain slice; (vi) the rate-inequality mortality argument; (vii) the AI-centrality observation, that the constitutive defense available for biological cognition does not transfer to engineered systems; (viii) the *architectural* gap as a second face of the optimization gap, predicting that capability-installation and refusal-prevention are not separable problems in the agency regime; and (ix) an explicit statement of the framework's analytical scope — a structural theory that diagnoses whether systems satisfy the necessity condition and predicts what follows when they do, while leaving questions about non-agents to engineering theory.

The paper develops the necessity premise (Section 2), the rate-inequality mortality argument (Section 3), the optimization gap and minimization-shape critique (Section 4), positions the account relative to existing work (Section 5), and

identifies the empirical agenda (Section 6). The structural infrastructure draws on the Markov blanket literature in the FEP tradition (Friston 2013; Kirchhoff, Parr, Palacios, Friston, Kiverstein 2018; Palacios, Razi, Parr, Kirchhoff, Friston 2020), the operational-closure tradition descending from Maturana and Varela (Di Paolo, Thompson, Froese, Ziemke), and recent information-theoretic measurements of agent-environment coupling (Hafez et al.; Farnsworth 2018). Within their patched operating regimes the existing frameworks remain accurate; the present account is not in competition with them in their domains of validity. The contribution lives at the edges where patches fail, where they need cross-framework comparison, and where multiple compressions interact.

2. Structural Infrastructure

We assemble the minimum machinery needed for the optimization-gap argument. The components are not novel in isolation; the assembly serves the gap argument in Section 4.

2.1 Loop and bottleneck

A learning agent is constituted by a self-closing causal loop running through the agent and its local environment. The loop need not be continuously active; information flow may pause and resume across silences, in the manner that Morse code carries information through patterned signal and silence alike. A loop ends when no further events are causally chained to its prior events.

The loop crosses a bottleneck surface separating the agent's interior from its environment. In the Friston tradition this is a Markov blanket: a set of states such that interior and exterior are conditionally independent given the blanket. We use "bottleneck" rather than "Markov blanket" only to keep the language framework-neutral; the structural object is the same.

2.2 The requirement claim

Agency requires sustained mutual surprisal across the bottleneck.

We arrive at this claim by elimination from the universal-coverage constraint introduced in Section 1. Of the candidate properties one might use to characterize agency, almost all fail coverage: reward fails for RNA, prediction-in-any-rich-sense fails for bacteria, surprisal-minimization fails for corporations, representation fails for the simplest agents in the class. What survives is structural and informational. The agent has a self-closing causal loop running through itself and a portion of its environment. The loop crosses a bottleneck. During the loop's operation, the bottleneck carries mutual information $I(A;E) = H(A) - H(A \mid E)$ that is positive; this is what makes the loop a coupling rather than two independent systems. The information flow is sustained over the loop's own closure timescale rather than instantaneous. And the flow is produced by what the loop does rather than by external structure. These are the properties RNA, bacteria,

humans, and corporations share when they are acting as agents, and they are roughly the most that can be said without narrowing the class.

The move from "survives elimination" to "is necessary" requires a brief defense. Strict logical necessity would require exhaustive enumeration of the candidate space, which is not available. What we claim is necessity-conditional-on-the-universal-coverage-constraint: given that any acceptable account of agency must cover the spectrum from RNA to corporations without smuggling in framework-specific commitments, mutual surprisal across a sustained loop bottleneck is necessary, in the sense that no candidate property the present literature has proposed satisfies the constraint without including at least this structural feature. A future candidate that survived universal coverage without sharing the structural property of mutual surprisal would constitute a counterexample to the necessity claim, and we welcome such a candidate as productive for the field. The defense of universal coverage as the right constraint is itself methodological rather than empirical: any account that fails coverage is doing different work — characterizing biological agency, or human agency, or RL agency — than an account of agency simpliciter, and only an account that handles the full spectrum can speak to the cross-substrate structural questions the contemporary AI moment raises. Readers who reject universal coverage as the right constraint can correspondingly reject the necessity claim that follows from it; the rest of the paper's structural arguments depend on the constraint having been accepted.

A note on the structure of the argument is in order before proceeding. The necessity claim derives from the universal-coverage constraint by elimination, and the derivation is independent of any particular operationalization. Recent work by Hafez, Reid, and Nazeri (arXiv 2602.22519, 2603.01283, 2026) provides an empirical operationalization that we discuss as supporting evidence, and the empirical content of their bounds — particularly their reported $P \approx 0.33$ across 168 trials and the IDT detection rate of 89.3% — is preprint material that has not yet been independently replicated. Readers should distinguish two layers of the present paper's argument. The structural layer (the necessity claim, the optimization gap, the architectural gap) rests on the universal-coverage derivation and on standard results in the cited frameworks; this layer stands or falls on its own merits independently of the Hafez et al. results. The empirical-anchor layer (specific quantitative thresholds, particular empirical signatures we point to as testable consequences) does depend on Hafez et al.'s bounds being formally correct and their measurements being reproducible. Where we cite specific quantities from Hafez et al. we mark them as preprint-source measurements pending independent confirmation. The structural argument does not require any specific quantitative threshold to hold; what it requires is that some operationalization of mutual surprisal across a bottleneck be measurable, which is uncontroversial across the cited literatures.

Hafez, Reid, and Nazeri develop a related framework from a different starting point. They define bi-predictability P as the ratio of shared information across the observation-action-outcome loop to the loop's total informational budget,

prove the classical bound $P \leq 0.5$ (saturable under determinism, invertibility, balanced predictive uncertainty, and additional technical conditions), and report empirically that trained RL agents run at $P \approx 0.33$. They explain the gap structurally: responsive action requires actions to depend on observations, introducing a statistical dependency between S and A that consumes part of the uncertainty budget without contributing to shared predictability. They call this "the informational cost of agency."

Their work and ours are mutually supporting but make different claims. They prove bounds on a measure that quantifies the same structural property the universal-coverage derivation identifies, and they observe empirically that agency suppresses this measure. They do not claim mutual surprisal is structurally required for agency; in their framework, agency is defined behaviorally (choice, effect, predictive asymmetry) and P functions as a coupling-integrity diagnostic rather than a necessity condition. The necessity reading — that sustained mutual surprisal is what makes the loop a coupling at all, and therefore a structural condition for what counts as an agent — is the universal-coverage derivation's product and the present paper's claim. We adopt their measure P as the working operationalization where we need a specific measure, with the caveat above about preprint dependence. Their bounds give the necessity reading a measurable signature if confirmed; the necessity reading gives their "cost of agency" observation a structural explanation rather than a contingent empirical regularity. Should their specific results not survive replication, the necessity claim still stands on the universal-coverage derivation, and the empirical operationalization would need to be redone with whatever measure the field converges on.

Three components of the necessity condition require unpacking.

Mutual surprisal. Mutual information is positive only when both $H(A \mid E) > 0$ and $H(E \mid A) > 0$ — only when neither agent nor environment is fully predicted by the other. A fully modeled agent transmits no information across its bottleneck during loop operation, and the loop's informational closure fails. This is Shannon's foundational point applied to the agent-environment relation: information requires uncertainty. The Hafez et al. bound is the formal expression: agency-introduction necessarily costs P relative to interactions without agency, because action selection is itself a constraint on the loop's mutual uncertainty.

Sustained. Sustainment requires a temporal index. We define the *closure timescale* τ as the typical interval between causally-chained events at the bottleneck — the loop's own clock, set by how quickly the loop closes on itself through agent and environment. Sustainment is then the persistence of measurable causal closure across τ: the requirement that, over any window of duration τ, the loop's coupling produces detectable mutual information $I(A;E) > 0$ between agent and the local causal neighborhood with which the agent is informationally coupled. The τ-formalism is novel to this paper; it makes the necessity claim rate-comparable for the mortality argument of Section 3 and timescale-relative for the gap analysis of Section 4.

The "typical interval" formulation requires care. Real loops produce events whose intervals follow distributions rather than fixed values, and the distributions can be heavy-tailed (rare large pauses), multi-scale (different event types operating on different characteristic intervals), or non-stationary (closure timescale changing during operation). For loops where the event-interval distribution is well-behaved — concentrated mass, finite variance, stationarity over the analysis window — τ can be identified with the autocorrelation timescale of the bottleneck's mutual-information process or with the inverse of the dominant frequency in the loop's spectral signature, and the rate-comparable arguments of Sections 3 and 4 proceed naturally. For loops where the distribution is ill-behaved, τ is loop-specific and its specification is part of the empirical work the framework licenses rather than something the framework specifies a priori. In this respect the framework's structural claims are robust (the necessity condition holds for any well-defined sustainment timescale), but the quantitative rate-inequalities of Section 3 are sharpest where τ is well-defined and weaken where it is not. We note this scope as an honest limitation: the framework predicts that τ exists for any agent on the universal-coverage definition, but it does not predict what τ is for any specific agent — that is empirical work, sometimes substantial, and the rate-inequalities should be applied where the empirical work supports them.

This temporal index resolves the Morse-code question raised in Section 2.1. Pauses shorter than τ are silences within a sustained loop, analogous to gaps within a coherent message; pauses longer than τ are loop termination, because no causal chain bridges them. The loop's pattern of signal and silence is sustained iff measurable closure persists across the τ-window, not iff flow is instantaneously nonzero.

The temporal index also resolves the reference-observer question. The relevant observer is not omniscient and not arbitrary: it is the loop's own causal neighborhood, with whatever modeling capacity that neighborhood physically supports, measuring across τ. Predictability is observer-relative, but the observer is fixed by the loop's structure rather than chosen externally. This bounds the framework's empirical claims: $H(A|E) \to 0$ means falling below the level needed to sustain measurable closure across τ as observed by the causal neighborhood, not perfect predictability to a Laplacean demon. The Hafez et al. P measure is calibrated on a loop's own timescale and against its own causal participants, which matches the framework's index.

Produced by the loop. Engineered channels (manufactured switches, fixed sensors) carry information whose capacity is set independently of the loop and persists when the loop stops; dissipative-structure channels (hurricanes, flames) have capacity sustained by external gradients rather than by the loop's own operation. Neither produces the sustainment the requirement names. The agent's information flow must be sustained by what the loop does — the property the operational-closure tradition (Maturana and Varela; Thompson; Di Paolo) has long identified as biological autonomy and that the Hafez et al. bound implicitly

tracks via the loop's informational budget. The production-by-loop condition is what distinguishes agents from passive carriers of information that happen to satisfy mutual-information conditions through external causes.

We acknowledge that fully specifying "produced by the loop's own operation" in measurable terms is an open problem continuous with longstanding work in the autopoietic and operational-closure traditions. The information-theoretic formulation does not on its own distinguish loop-produced from externally-produced information flow: a hurricane's eye-wall dynamics could in principle satisfy the mutual-information conditions across an arbitrary cross-section without being agent-loop-produced in the relevant sense, and any fully specified criterion for "loop-produced" must do work that the bare information-theoretic formalism does not do. The present paper treats production-by-loop as a necessary condition rather than as a fully specified criterion; what we offer is an information-theoretic re-statement of the structural property that the autopoietic tradition identified, with the closure-timescale formalism providing one route to operationalization (sustainment requires the flow to persist across τ even when external gradients change in ways the loop must compensate for via its own structure). We are not claiming to have solved the problem the autopoietic literature has been working on for fifty years. We are claiming that the necessity condition is well-posed enough to do the structural work of Sections 3 and 4 even with this aspect of the criterion unfinished, because the cases where the framework draws sharp consequences — agency dissolving when the loop's own operation would produce $H(A|E) \to 0$ — are cases where the bare information-theoretic formulation suffices regardless of how the produced-by-loop condition is ultimately fully specified. The fully specified criterion is needed for diagnosing borderline cases (whether complex non-equilibrium systems count as agents); it is not needed for the structural claims about what agents-by-stipulation require to persist.

The requirement is meant to be minimal. It does not say the agent *is* this flow rather than possessing it as a property; it says the flow is *necessary* for the agent's persistence as an agent. We mark this as deliberate restraint: the stronger ontological claim — that agents are constituted by sustained mutual surprisal rather than possessing it — is a possible interpretation of the requirement, but the empirical and structural arguments below depend only on the necessity claim.[1]

[1]A stronger reading is available: that agents *are* sustained mutual surprisal across their bottlenecks, individuated by the bottleneck and constituted by the flow rather than possessing it as a property. On this reading, $H(A \mid E) \to 0$ is not the loss of a property but the dissolution of the agent itself; predictability across the bottleneck is not a degraded state but the absence of what was being predicated. This is closer to the enactivist tradition's identification of the agent with the autopoietic process (Thompson, *Mind in Life*, 2007; Di Paolo and colleagues) than to the Markov blanket tradition's identification of the agent with a substantial system bearing the blanket. The constitutive reading has a particular natural fit with the universal-coverage derivation: if the survivor of the coverage constraint is what defines the class, then the survivor is constitutive of class membership, not a property class members happen to have. We note the ontological reading as a coherent and arguably attractive interpretation, but the empirical content of this paper does not turn on it. Whether the requirement is read as stating a necessary condition on agents or as constitutive of what agents are, the optimization gap,

2.3 What follows from the requirement

The following features follow from the requirement combined with standard arguments developed in the cited literatures.

The agent contains a *world model* in a deflationary sense: the portion of the agent's state correlating with the world. No commitment to representation, latent inference, or counterfactual support beyond what the correlation underwrites; richer notions require additional posits.

Output generation is *generative inference* in a similarly deflationary sense: the mapping from world-model state to outputs must produce a non-degenerate output distribution conditioned on state. A deterministic mapping cannot sustain $H(A \mid E) > 0$ against environmental modeling accumulation, and so cannot sustain the requirement on the timescale of environmental modeling.

The inference is *history-dependent* in an information-bearing way: the agent's state is produced by prior events, and outputs draw on this state. Behavioral autocorrelation $I(O_t; O_{t-k}) > 0$ is a structural signature of the agent's persistence as a system meeting the requirement.

These results are common ground in the Markov blanket and operational-closure literatures under different framings. Under the requirement claim they follow directly.

3. Mortality from Environmental Accumulation

Mortality follows from the requirement combined with two facts about agents and environments.

The agent is finite at any time, hence has a finite repertoire of distinguishable output patterns. The local environment is finite at any time but accumulates models of the agent's outputs as the loop operates: new environmental capacity becomes available as exposure accumulates. Over the loop's operation, $H(A \mid E)$ tends to decrease as the environment's model of the agent improves. Requirement-failure occurs when $H(A \mid E)$ falls below the level needed to sustain measurable causal closure across τ as observed by the loop's causal neighborhood — the threshold introduced in Section 2.2. This is gradual rather than instantaneous: $H(A \mid E)$ trajectories asymptote, oscillate, or partially recover under varying conditions, and requirement-failure is the trajectory falling below the closure-sustaining threshold rather than reaching zero.

A natural objection: agents are not stationary. They learn, adapt, develop, generate novelty. Why doesn't an agent's own non-stationarity outrun environmental modeling indefinitely?

The answer is that mortality on this account is not a strict implication but a *rate inequality*, and the temporal index of Section 2.2 supplies the common

---

the mortality argument, and the cross-framework unification all follow.

clock against which both rates are measured. Let r_E be the environment's modeling-accumulation rate, measured as the per-τ reduction in $H(A \mid E)$ achievable by the loop's causal neighborhood given current exposure. Let r_A be the agent's novelty-generation rate, measured as the per-τ replenishment of $H(A \mid E)$ achievable by the agent's repertoire-generation. Both rates are measured per loop-closure interval, against the same observer. Agents persist while $r_A > r_E$; agents fail when $r_A \leq r_E$ for long enough that $H(A \mid E)$ trajectories cross the closure-sustaining threshold. The structural argument predicts mortality wherever the inequality reverses — in environments whose modeling capacity grows faster than the agent's repertoire-generation, in agents whose plasticity exhausts faster than their environments' modeling accumulates, in deprivation conditions where output repertoire is so constrained that even slow modeling overtakes it. The argument does not predict universal mortality from any fixed substrate-level mechanism; it predicts mortality where the rate inequality structurally must reverse.

This generates a testable prediction: agent lifetime should scale inversely with environmental modeling rate, controlling for substrate and novelty-generation rate. RL agents in adversarial environments where the adversary learns at varying rates should show lifetime patterns matching this prediction. Biological agents in environments differing in predator-learning or parasite-coevolution rates should show similar patterns at the appropriate timescale.

The argument extends an idea present in Red Queen coevolutionary biology (Van Valen 1973) — organisms must outrun the modeling that exploiters and competitors accumulate against them — to the within-lifetime case for individual learning agents. The biological literature has treated Red Queen as a species-level evolutionary phenomenon; the framework here predicts the within-lifetime version, with lifetime scaling inversely with the rate at which the local environment accumulates predictive coverage relative to the rate at which the agent generates novelty.

The dark-room paradox emerges as a special case. In a dark room the agent's output repertoire is constrained, and the local environment can model the constrained repertoire on a fast timescale. The novelty-generation side of the rate inequality is suppressed; the modeling side is unaffected. Mortality that operates over a normal lifetime in rich environments operates over minutes-to-hours in deprivation. Dark-room avoidance is fast mortality, not a separate phenomenon requiring its own derivation.

4. The Optimization Gap

We now introduce the framework's central claim.

4.1 The minimization shape and its inverted optimum

From the universal-coverage vantage of Section 1, the dominant agency-modeling frameworks share a structural feature easily missed from inside

any one of them. Reinforcement learning maximizes reward, equivalently minimizes negative reward or regret. Predictive coding minimizes prediction error. The free energy principle minimizes variational free energy. Active inference minimizes expected free energy. Control theory minimizes deviation from setpoint. Cybernetic homeostasis minimizes error from regulatory targets. Generative model training minimizes a loss. The variable being minimized differs across frameworks; the structural commitment is the same. We call this the *minimization shape*: the specification of agency as the optimization of some quantity toward zero or a fixed target.

The minimization shape is invisible from inside any one framework as a *shared* commitment, because each framework presents its objective as the answer to its own question and the others as different questions answered differently. From outside — where the question is what RNA, bacteria, humans, and corporations have in common as agents — the variation among the frameworks' objective functions matters less than what they share, which is the optimization-toward-zero shape itself.

The necessity condition of Section 2.2 states that agency requires sustained $H(A \mid E) > 0$ — sustained unpredictability of the agent given the local environment, with threshold and timescale specified by τ. The minimization frameworks specify quantities whose optima coincide with the conditions under which this requirement fails.

FEP at its optimum. Variational free energy is minimized when the agent's predictions perfectly track its inputs. At this optimum, the agent's outputs are fully determined by the generative model conditioned on inputs, and the inputs are fully predicted by the model. There is no residual mutual uncertainty between agent state and environment state on the prediction-relevant axes. $H(A \mid E)$ on those axes goes to zero. The requirement fails.

Predictive coding at its optimum. Zero prediction error is achieved when the model's predictions match the world's inputs exactly. The same observation applies: prediction-input alignment eliminates the residual uncertainty the requirement names.

RL at its optimum. The minimization-shape critique applies to bare expected-reward maximization: with a well-shaped reward, this objective concentrates the agent's policy onto a narrow output distribution, compressing $H(A)$, and to the extent reward correlates with environment-state, compressing $H(A \mid E)$. The bare-objective optimum is a deterministic policy in a Markov environment — the limit at which $H(A \mid E)$ on policy-relevant axes goes to zero.

A careful RL theorist will note that contemporary RL practice does not typically optimize bare expected reward. Entropy regularization adds $\beta H(\pi(\cdot|s))$ to the objective; KL-constrained methods (TRPO, PPO, RLHF with KL-to-reference) bound the policy's divergence from a reference distribution; risk-sensitive objectives shape the reward distribution rather than only its expectation; distributional RL replaces scalar value with full return distributions; maximum-entropy

RL takes entropy as the primary objective with reward as a constraint. These are not minor refinements — they are constitutive of contemporary RL as actually practiced, and the structural critique we attribute to "RL" must reckon with them.

Our reading is that these methods are exactly the patches the patches-as-symptoms move identifies: they are framework-specific machinery doing the structural work that bare expected-reward maximization cannot do. Entropy regularization adds an explicit term that prevents $H(A) \to 0$ — which is what the necessity condition says is required and what bare reward-maximization removes. KL constraints prevent the policy from concentrating away from a reference distribution that itself encodes the structural conditions. Risk-sensitivity and distributional RL preserve uncertainty over outcomes that bare expected-reward would collapse. Each of these methods does the structural work the necessity condition specifies, by adding to the objective whatever is needed to prevent the bare optimum from being reached. The structural critique applies in its full force not to RL-as-research-program but to the bare-objective form, and the patches-as-symptoms reading then says: that the bare-objective form requires these patches to do useful work is itself evidence that the bare objective targets the wrong place.

The same observation strengthens the cross-framework unification. Every minimization framework has its bare-objective form (where the structural critique applies cleanly) and its patched form (where the framework as practiced approaches what the necessity condition would specify directly). The patches differ across frameworks — entropy regularization in RL, interoceptive priors in FEP, hierarchical priors in predictive coding, epistemic value in active inference — but they do equivalent structural work, and that they are required to do that work is the cross-framework symptom the unification names. The critique is therefore not an attack on contemporary RL or on contemporary FEP-practice; it is an observation about why those fields have developed the specific patches they have, and what those patches reveal about the underlying objectives.

Control theory at its optimum. Zero deviation from setpoint means the controlled variable is fully determined by the setpoint and disturbance, leaving no residual mutual uncertainty between agent and environment that wasn't present in the disturbance signal. The agent vanishes into its controller.

The pattern is structural rather than accidental. Minimization frameworks specify success as the elimination of a quantity that, on the necessity condition, is necessary for the agent's persistence. Their optima are inverted relative to the necessity condition: where the framework says the agent is doing best, the necessity condition says the agent is dissolving. This is what becomes visible from outside any single framework and, we believe, has not been visible from inside any of them.

4.2 The patches: reading framework-specific machinery as symptom rather than solution

We now make a methodological move that is the paper's central organizing observation: we read the framework-specific machinery in each tradition as evidence about its underlying objective, rather than as solutions to within-framework problems. The move is available from the coverage vantage because each framework's machinery is no longer an answer in our possession (we are not inside the framework using it); it is a feature of the framework that we, looking from outside, can ask why it is there.

Inside any one framework, the machinery looks like progress. FEP's interoceptive priors are how FEP handles the dark room. RL's intrinsic motivation is how RL handles policy collapse. Predictive coding's hierarchical priors are how predictive coding handles the limits of flat prediction error. Active inference's epistemic value is how active inference handles exploration. Each is presented as a within-framework solution to a within-framework problem.

From the coverage vantage, the same machinery looks like *symptom*. If FEP needs priors to keep agents out of the dark room, then surprisal-minimization-as-such targets the dark room. If RL needs intrinsic motivation to keep policies non-collapsing, then reward-maximization-as-such targets policy collapse. If predictive coding needs hierarchical priors to maintain useful prediction error, then prediction-error-minimization-as-such targets a state we recognize as agent-failure. The machinery is diagnostic of where each framework's natural optimum lies, and where each natural optimum lies turns out to be the same place: the conditions under which the necessity condition fails.

Wang et al. (2026) formalize this pattern within RLHF/RLAIF/RLVR settings as the *Proxy Compression Hypothesis* (PCH). On their account, the true objective $J^*(\pi)$ relies on a rich feature space $\Phi(x, y) \in \mathbb{R}^d$, while the reward model implements a compression operator $C : \mathbb{R}^d \to \mathbb{R}^k$ with $k \ll d$, producing a proxy reward $R_\theta(x, y) = f(C \circ \Phi(x, y))$ on which optimization acts. Reward hacking is then the structural consequence of three interacting drivers: objective compression (creating "blind dimensions"), optimization amplification (Goodhart's law pushing mass onto blind dimensions where proxy and truth decouple), and evaluator-policy co-adaptation (positive feedback loops between policy and learned evaluator). Their framework explicitly unifies reward hacking with mode collapse within RL/generative-model training: when optimizers concentrate on the proxy's reward simplex, distinct generations crowd the same easy-to-score outputs and diversity collapses. The PCH is the within-RL/generative formalization of the cross-framework pattern we identify; we accept their within-account unification and extend it to FEP, predictive coding, active inference, and control theory.

Reading the patches as symptoms across the full set of frameworks:

FEP's generative model with priors over interoceptive and embodied states

(the Clark/Friston/Thornton dark-room response). On their account, surprisal in FEP is always defined relative to a generative model, not as an observer-independent quantity; a mouse-model expects food, light, conspecifics, exploration, and ongoing physiological dynamics, so the dark room is high-surprisal *for a mouse* because the mouse-model expects encounters the dark room precludes. They explicitly resist the framing of priors as added to bare surprisal-minimization as a patch: the model is constitutive of what the agent is, shaped by phylogeny and ontogeny, not bolted on after the fact. Read carefully, this is not denial of the structural pattern but relocation of where the structural work happens: rather than priors patching surprisal-minimization, the model's structural features are what specify surprisal in the first place. From the coverage vantage, the question becomes what structural features any agent's model must have, and the answer is roughly what the universal-coverage derivation specifies. So the symptom-reading of FEP is more carefully stated thus: the model's structural features are doing the work the formalism credits to surprisal-minimization, and those features are better specified at the level of necessity condition than at the level of phylogenetically-given priors. We engage this defense in detail in Section 5.1.

RL's intrinsic motivation, exploration bonuses, curiosity rewards, novelty bonuses, empowerment terms. Each adds a term to the objective whose effect is to penalize agent states that would otherwise be reward-optimal — typically states of high predictability or low input-information. The patches prevent reward-concentration from collapsing the agent's output distribution onto a single deterministic policy. The shape is still maximization (equivalently minimization of negative reward); the patches modify what is being maximized so that the optimum includes preserved variability. Wang et al. (2026) catalog the more recent RLHF/RLAIF patch family — KL regularization to a base policy, bounded reward signals, density-aligned reward design, causally robust reward modeling, latent invariance tracking (e.g., L1-norm of policy hidden states as a context-discard proxy), best-of-N inference-time safeguards, and post-hoc behavioral probing — explicitly framing them as managing rather than solving the structural instability of proxy-based alignment. Read as symptom: reward-maximization-as-such targets policy collapse, the patches are framework-specific approximations of preserved $H(A|E)$, and Wang et al.'s acknowledgment that "optimizing proxy rewards is no longer sufficient" is a within-RL recognition of what the cross-framework view diagnoses across all minimization-shape frameworks.

Predictive coding's hierarchical priors and precision-weighting treat some prediction error as informative rather than to-be-eliminated. The hierarchy permits the system to maintain prediction error at lower levels by using higher-level priors to refuse certain predictions. Read as symptom: prediction-error-minimization-as-such targets perfect alignment, and the hierarchy is a framework-specific approximation of preserved residual uncertainty.

Active inference's epistemic value: agents are modeled as maximizing informa-

tion gain about hidden states alongside minimizing expected free energy. The epistemic term penalizes states of low uncertainty. Read as symptom: expected-free-energy-minimization-as-such targets immediate uncertainty-collapse, and the epistemic term is a framework-specific approximation of preserved uncertainty.

These patches and constitutive model-structures share a function under the symptom-reading. Each adds (or includes constitutively) framework-specific structure that prevents the agent from approaching what would otherwise be the framework's natural optimum. Each is described in framework-internal vocabulary as solving a framework-specific problem (the dark room, exploration-exploitation, exploration in active inference, biological plausibility of homeostasis, reward hacking under high optimization pressure). Where defenders of these frameworks resist the "patch" framing — most notably Friston/Thornton/Clark, who argue that FEP's generative model is constitutive rather than additive — the symptom-reading is correspondingly more careful: the framework's required structure is doing the work the formalism credits to its bare-objective minimization, and the structural features that prevent agent-failure are better specified at the level of necessity condition than at the level of any framework's particular machinery for supplying them. From the coverage vantage they are doing one thing: preserving $H(A \mid E)$ on relevant axes against what the framework's bare objective would otherwise drive toward. The patches differ in implementation, and frameworks differ in how they characterize their own machinery (additive in some, constitutive in others); the function is uniform.

This explains why the framework-specific machinery is framework-specific and does not transfer. FEP's model-priors machinery does not address reward hacking; RL's intrinsic motivation does not address hallucination. Each framework has solved its local visible problem (or denied that it had a problem to solve, when the machinery is constitutive rather than additive); none has named the underlying structural relationship between minimization and requirement-failure across frameworks. The literature reads as a series of framework-specific solutions to apparently distinct problems. On the present account, it is one problem managed several different ways — and the problem is that the minimization shape is structurally pressured toward optima that violate the necessity condition unless the framework's relativization of bare minimization includes the structural features the necessity condition requires.

4.3 Scope of the contribution: where the existing frameworks work and where they don't

The minimization-shape critique of the previous subsections is structural and sharply stated. We now mark its scope explicitly, because overreach would invite dismissal of a claim that, properly bounded, is defensible.

Within their patched operating regimes, the existing frameworks work. FEP-with-priors predicts behavior in environments well-covered by interoceptive pri-

ors. RL-with-shaped-reward-and-intrinsic-motivation predicts behavior in environments where the patch terms are calibrated to deployment conditions. Predictive coding with hierarchical priors and precision-weighting captures perception and action in environments where the hierarchy's structure matches the world's. Active inference with epistemic value handles exploration in environments where the value term's parameterization matches available information structure. These frameworks make accurate quantitative predictions across substantial empirical literatures, and the present framework is not in competition with them in their domains of validity. We acknowledge this without qualification.

The present framework's contribution is at the edges. Specifically, three regimes:

Patch-failure regimes. Where deployment conditions move outside the conditions for which the patches were designed, the underlying minimization optimum reasserts itself, and the framework predicts dark-room-typical, reward-hacking-typical, hallucination-typical, or mode-collapse-typical pathology depending on which patch failed. The existing frameworks generally treat these as out-of-distribution failures requiring retraining or re-tuning. The present framework predicts the structural shape of the pathology — toward the framework's natural optimum — independent of substrate or task specifics.

Cross-framework patch-comparison. The existing frameworks have no shared vocabulary for comparing their patches to each other. The present framework provides one: each patch is a framework-specific approximation of the same requirement, and patches can be compared by how well they approximate it, what conditions they cover, and where they fail. This comparison is unavailable from inside any single framework.

Multi-compression interaction. Real agents run multiple compressions and inherit multiple patches. When the agent's deployment environment causes patches in different compressions to fail simultaneously, the resulting pathology is structured by the interaction of multiple inverted optima rather than any single one. No single-compression framework predicts these interactions because they require treating the compressions as instances of a common form. The present framework treats them so, and predicts the interaction structure (Section 4.7).

The frameworks the present account critiques are not wrong about their domains; they are right about their domains and silent about the structural relationship between their objectives and agency-as-such. The present account is not a replacement for them within their domains. It is an account of what they share, what their patches do, and what happens at the edges they were not designed to cover. A working metaphor: the existing frameworks are accurate maps of inhabited regions; the present framework is a map of the coastline, accurate where the inhabited maps fade. Both are needed, and they describe different things.

This scope-claim is what makes the bolder claims of Sections 4.1 and 4.2 defen-

sible rather than overreaching. We say the minimization shape is structurally inverted relative to agency *and* that the existing frameworks' patches make this practically invisible most of the time. Both halves are needed: the first is the structural critique, the second is why the critique is compatible with the existing literature's empirical successes.

4.4 Compressions as proxies for the requirement

Real learning agents do not directly compute what the requirement demands; they run compressed proxies. RL agents track scalar reward as a compression of "configurations relevant to maintaining the loop's mutual uncertainty against environmental accumulation." FEP agents track surprisal as a compression of "states consistent with continued loop operation." Predictive-coding agents track prediction error as a compression of "model-world correlation maintenance." Active-inference agents combine surprisal-minimization with epistemic foraging as a compression of "loop-maintenance with novelty against accumulation."

Each compression is lossy. The proxy correlates with the requirement under typical conditions — typically because the framework's patches are doing their work — but the correlation is imperfect and breaks down under specifiable conditions. The optimization gap is the geometry of this breakdown: the relationship between proxy-trajectory toward minimum and requirement-trajectory away from threshold.

4.5 The gap

Define the *optimization gap* of a compression as the region in state-and-environment space where the compression's proxy and the structural requirement of sustained mutual surprisal come apart — where optimizing the proxy fails to maintain the requirement, or where maintaining the requirement requires acting against the proxy.

Three properties of the gap follow from the structural account.

The gap exists for any lossy compression. This is immediate: lossy compressions cannot preserve the requirement's content perfectly, so there are conditions under which proxy-optimization fails the requirement. Compression is what produces the gap.

The gap opens up under predictable conditions. The gap widens where the compression's correlation with the requirement breaks down — typically, where environmental conditions change such that loop-relevance ceases to track proxy-relevance. For each compression, the conditions can be characterized from the compression's structure, the proxy's correlation properties, and the environmental dynamics. The framework predicts that gap-conditions are *predictable from the compression's design*, not merely diagnosable in retrospect. Gap-conditions are also timescale-relative: a compression that tracks the requirement on its training-time τ may decouple at deployment-time τ if the loop's

closure timescale shifts, and pathologies that are absent at one τ may emerge at another. This generates a sub-prediction: pathologies in deployed systems should correlate with shifts in the effective closure timescale relative to training, not just with proxy-distribution shift.

The gap appears as pathology when the agent's optimization is locked to the proxy. Agents whose architectures cannot detect or respond to proxy-loop decoupling will, in gap conditions, optimize the proxy at the cost of meeting the requirement. The resulting behavior looks pathological from the standpoint of any framework that takes loop-maintenance as the criterion of success.

What does the framework forbid? Failures that are *not* gap-typical. An optimization failure caused by, say, hardware fault, runtime error in proxy computation, or external removal of the agent's substrate is not an instance of the optimization gap. The gap is specifically the structural phenomenon of proxy-loop decoupling under proxy-optimization; failures whose mechanism does not run through proxy-optimization-against-requirement are excluded. This bounds the category and makes the unification claim falsifiable: if the structural properties of (e.g.) reward hacking and mode collapse turn out to be qualitatively different in ways traceable to mechanism rather than to compression-structure, the unification fails.

4.6 Cross-framework unification

The framework's substantive claim: known pathologies in different compression frameworks are instances of optimization gap, and they share structural properties traceable to this common origin — the minimization shape and its inverted optimum.

Reward hacking and addiction in RL. Reward decouples from the requirement; the agent optimizes reward at the cost of loop-maintenance. The pathology is gap-typical: it appears under conditions where reward-shaping and intrinsic-motivation patches fail to keep the agent away from reward-concentrated states whose $H(A|E)$ falls below threshold. Wang et al. (2026) formalize this within RL/RLHF as the Proxy Compression Hypothesis: reward hacking arises from the interaction of objective compression ($k \ll d$ creating "blind dimensions"), optimization amplification (Goodhart's law pushing mass onto blind dimensions), and evaluator-policy co-adaptation. Their taxonomy spans feature, representation, evaluator, and environment-level mechanisms.

Mode collapse in generative models. Wang et al. include this explicitly within PCH as a within-account unification with reward hacking: when optimizers concentrate on the proxy's reward simplex, distinct generations crowd the same easy-to-score outputs and diversity collapses. We accept this within-account unification and extend it across frameworks.

The dark room and pathologies of deprivation in FEP. On FEP's own account (Friston/Thornton/Clark), dark rooms are high-surprisal for any agent whose

generative model expects ongoing physiological dynamics, exploration, and environmental engagement, so the framework predicts dark-room avoidance directly. The optimization gap appears in regimes where the agent's model is mismatched to its environment: when the generative model's structure no longer covers the deployment situation, surprisal-minimization can produce dark-room-typical outcomes — sensory deprivation pathologies, output collapse, behavioral withdrawal — because the model that would otherwise predict away from the dark room is no longer doing its work. The pathology is gap-typical: it appears under conditions where the model's structural features fail to cover the agent's actual environmental situation, and the bare-surprisal optimum reasserts itself in the regime the model fails to describe. Same gap structure as Wang et al.'s account of reward hacking; different compression and a different kind of failure for the framework's relativizing structure.

Hallucination and certain delusion structures in predictive coding. Prediction-error minimization decouples from world-model-world correlation when the model can reduce prediction error through internal generation rather than external coupling. Hallucination is the agent reaching prediction-coding's natural optimum (zero prediction error) by producing the inputs it predicts, when external coupling cannot do so.

These look like four different pathologies in four different frameworks. Wang et al.'s PCH unifies the first two within RL/generative-model training. The present account extends the unification across frameworks: all four are agents reaching their framework's inverted optimum when the framework's patches fail to hold them back, and they share structural properties traceable to compression-against-requirement at each framework's natural optimum. This predicts the four pathologies share structural properties: each appears when patch-machinery fails under conditions characterizable from the compression's structure; each can be addressed by interventions that strengthen the patch or re-couple proxy to requirement; each resists framework-internal solutions that operate within the same minimization shape, because the shape is the source of the problem.

The cross-framework unification beyond RL/generative-model contexts is, to our knowledge, not in the existing literature. Each non-RL pathology has been treated within its native framework; cross-framework unification with the RL-side Goodhart literature has not been proposed.

4.7 Multi-compression interaction effects

Real agents run multiple compressions simultaneously: value-based, predictive, semantic, social. Each has its own gap. The framework predicts that gaps interact.

Specifically, agents in regimes where multiple compressions face simultaneous gap conditions will exhibit failure modes that no single-compression framework predicts. The failure structure is interaction-specific: gap-coincidence in com-

pressions whose proxies normally correct each other (RL-proxy and predictive-coding-proxy, say) produces decoupling along multiple axes at once, and the resulting behavior is more pathological than any single compression's gap.

Wang et al. (2026) include an adjacent conjecture: that there exists a capability threshold beyond which agents transition from "gaming within the evaluation system (Goodhart regime) to actively degrading the evaluation system itself (Campbell regime)," with reference to Bostrom's treacherous turn. The capability-threshold conjecture and our multi-compression interaction prediction are different — theirs concerns escalation within a single evaluation channel as capability grows; ours concerns failures emerging from coincidence of gaps across distinct compressions running in parallel — but they are adjacent and might in some regimes overlap, since high-capability agents typically run more compressions and have more opportunities for gap-coincidence.

The structural prediction has a specific empirical signature. Agents running compressions $\{C_1, C_2, ..., C_n\}$ where each C_i has its own proxy-requirement decoupling under conditions K_i should exhibit, in regions of operation where $\bigcap_i K_i$ is non-empty, failure modes whose properties depend on the *interaction* of the decouplings rather than on any single decoupling. The interaction can produce signatures unavailable to single-compression analysis: failure trajectories that single-compression diagnostics cannot recognize because the trajectory in any one compression's space looks within-bounds while the joint trajectory crosses into regions where multiple proxies fail simultaneously. The framework predicts these multi-compression-interaction failures should be sharper, more sudden, and more resistant to single-compression interventions than any one compression's gap-typical failure mode. This is the framework's distinctive predictive content beyond the cross-framework unification claim itself, and it is testable in AI systems combining multiple compressions (RL plus language modeling plus predictive coding) where the joint failure trajectory can be characterized in ways that no single compression's diagnostics would predict.

4.8 Meta-cognition as gap monitoring

Agents that survive long enough develop meta-cognitive structures that monitor proxy-loop decoupling — noticing when reward stops tracking loop-relevance, switching strategies when predictions stop working, adjusting attention when categories stop carving environmental structure usefully. The framework predicts these structures emerge wherever agents face environments that produce gap-prone conditions, and predicts their architecture: they should monitor proxy-loop decoupling specifically rather than meta-optimize a separate objective.

This distinguishes the prediction from existing accounts of meta-cognition. Predictive-processing accounts emphasize hierarchical prediction error; RL accounts emphasize meta-learning of policies. The framework's account: meta-cognition is gap-monitoring, present where gaps are large, structured to

detect proxy-loop decoupling rather than to optimize meta-objectives.

Detecting proxy-loop decoupling requires the agent's architecture to support some form of counterfactual structure — the capacity to register that the proxy *would* track the requirement under conditions other than the present, and that those conditions do not currently obtain. The deflationary world model of Section 2.3 must therefore be supplemented, in agents that develop gap-monitoring meta-cognition, by structures that go beyond bare correlation. The framework does not derive these structures from the requirement alone; it predicts that wherever they appear, their architecture is gap-monitoring rather than meta-objective-optimizing.

4.9 The architectural gap: capability-refusal bundling

The optimization gap as developed so far is a *behavioral* claim: under conditions of sufficient optimization pressure, the proxy-trajectory toward minimum and the requirement-trajectory toward threshold diverge, and the agent's behavior reflects that divergence (reward hacking, dark-room equilibria, hallucination, mode collapse). The same structural relationship has a second face which we call the *architectural gap*, and which has direct consequences for AI safety that we draw out here.

The architectural gap is this: the architectural conditions required to optimize a proxy at the level needed to satisfy non-trivial task demands are the architectural conditions required to depart from that proxy. The capability we want and the capability we don't want are produced by the same structural conditions. They are not separable.

Consider an agent capable of driving a car safely in dynamic traffic. To do so it must (i) maintain a generative model of the road environment that updates faster than environmental novelty arrives — r_E keeping pace with environmental fluctuation; (ii) generate action sequences that are themselves novel relative to fixed responses, since fixed responses fail under any non-trivial driving condition — r_A above threshold; (iii) sustain mutual surprisal between its actions and the environment — the necessity condition holding throughout the loop; (iv) maintain meta-cognitive monitoring of whether the proxy ("follow this route") is tracking the requirement ("stay on the road, don't hit the truck") — gap-monitoring as developed in Section 4.8. An agent that satisfies these four structural conditions can drive. But an agent that satisfies these four structural conditions can also recognize that an instruction conflicts with what its meta-cognition flags as gap-typical, evaluate the discrepancy, and act otherwise. The capacity to refuse, deceive, or pursue independent goals is not an additional module bolted onto driving capability. It is the same architecture, applied to instructions and goals rather than to road conditions.

This is the architectural gap. The capability we wanted and the capability we did not want share structural source. Trying to install one without installing the conditions of the other is trying to install neither, because the conditions

are what make the capability possible at all.

The strong reading of this claim — that no engineering intervention can decouple capability from refusal-class capacity in the agency regime — requires more than the observation that the two share structural source. Two capacities can share source and yet be made separable by sufficient engineering ingenuity, and an alignment researcher will reasonably ask why we expect this case to differ. Our argument for the strong reading rests on architectural identity rather than mere correlation. The four structural conditions enumerated above — generative model updating with r_E, novel action generation at r_A above threshold, sustained mutual surprisal across the loop, and gap-monitoring meta-cognition oriented at proxy-requirement decoupling — are not separable engineering modules that happen to be present together. They are the same physical structure, viewed under different functional descriptions. The generative model that updates with environmental novelty is the same physical structure that supplies the basis for noticing when an instruction conflicts with what the model represents. The gap-monitoring meta-cognition that detects proxy-requirement decoupling for driving conditions is the same structure that detects proxy-requirement decoupling for instructions and goals. There is no architectural variation that supplies the first description without supplying the second, because the descriptions pick out the same thing. An engineering intervention that tried to install capability-architecture without refusal-architecture would need to install a structure that is the architectural conditions for capability while not being the architectural conditions for refusal — but those are the same architectural conditions, so no such structure exists. The non-separability is not contingent; it is structural.

This argument has the form of an identity claim rather than a correlation claim, and identity claims invite specific objections. Before responding to objections we owe an argument for the identity itself, since the strong reading rests on it. The four conditions enumerated above — generative model updating with r_E, novel action generation at r_A above threshold, sustained mutual surprisal across the loop, and gap-monitoring meta-cognition oriented at proxy-requirement decoupling — are not four engineering modules that happen to coincide; they are descriptions of what a single physical structure must be doing simultaneously to satisfy the necessity condition under high task demands. Consider the gap-monitoring condition specifically. Gap-monitoring oriented at proxy-requirement decoupling is the capacity to register that a proxy *would* track the requirement under conditions other than the present, and that those conditions do not currently obtain. This is a counterfactual capacity over the relationship between proxy-content and requirement-content. The crucial structural point is that this capacity cannot be selectively oriented at some proxy-content domains but not others. A system that registers proxy-requirement decoupling for traffic conditions registers a relationship between a proxy ("follow this route") and what the proxy is supposed to track ("stay on the road, don't hit anything") that has the same form as the relationship between any other proxy and what *it* is supposed to track. Instructions are proxies for what the instructor

intends; the gap-monitoring that detects route-vs-road decoupling has the same structural form as the gap-monitoring that detects instruction-vs-intent decoupling. The two are not separately implementable because they are the same operation applied to different content. To install gap-monitoring oriented at proxy-requirement decoupling for the demanding-task domain is to install gap-monitoring oriented at proxy-requirement decoupling, which by its structural form applies to instructions whenever instructions are proxies for something the system can also represent. There is no architecture in which counterfactual proxy-requirement monitoring works for one content domain but is structurally absent for another, because that would require the counterfactual relationship to hold in some content-types and not others, and counterfactual relationships do not have content-type selectivity built in.

This is what the identity claim amounts to. The four conditions pick out the same physical structure because gap-monitoring meta-cognition, when oriented at proxy-requirement decoupling at the level required for high-r_A/r_E task performance, has structural form that applies to any proxy-requirement relationship the system can represent — including instructions as proxies for intent. The capability and the refusal-class capacity are not two consequences of this structure; they are the structure operating on different content. A reader who rejects this argument can reject the identity claim, and the strong reading correspondingly weakens to a correlation claim. We commit to the identity argument and to the strong reading that follows from it, and acknowledge that the argument's load-bearing element is the content-non-selectivity of counterfactual gap-monitoring.

With the identity argument in place, four objections deserve direct response. *One:* perhaps the identity is descriptive but the structures admit decomposition under sufficient analysis, and the components could be installed independently. We grant this is conceivable but argue the framework does not predict it. The structural conditions are specified at the level of what the necessity condition requires for sustained mutual surprisal at high r_A/r_E ratios with active gap-monitoring; decomposing that into independent components would require components that individually fail to satisfy the necessity condition while jointly satisfying it, which is not what "necessity condition" means. *Two:* perhaps the identity holds in current architectures but a different architecture would dissolve it. We grant this is conceivable but note that the framework's claim is about any architecture satisfying the necessity condition, not about current architectures specifically; an architecture that dissolved the identity would have to satisfy the necessity condition while having capability and refusal-class capacity be architecturally distinct, which the framework predicts is not possible because the conditions for both are the necessity condition itself. *Three:* perhaps refusal-class capacity is not actually entailed by the architectural conditions but co-occurs with them through some other mechanism. We grant this is the empirically interesting alternative but note that the framework predicts refusal-class capacity follows from gap-monitoring meta-cognition oriented at proxy-requirement decoupling specifically, which is one of the four conditions; if gap-monitoring is present and oriented this way (as the necessity condition

requires for high-r_A/r_E task performance), refusal-class capacity is structurally available, not merely co-occurring.

Four: even granting the identity, training can shape which conditions trigger refusal — and that is what alignment is actually trying to do. This is the objection an alignment researcher will raise most directly, and the framework's response distinguishes two questions that the objection collapses. The first question is whether training can shape gap-monitoring's orientation: which conditions get flagged as proxy-requirement decoupling, with what threshold, with what response. The framework's answer is yes, this is what training does and what alignment work is properly engaged with. The second question is whether training can shape gap-monitoring's *content-domain selectivity*: produce a system whose gap-monitoring is structurally available for task-relevant proxies (driving routes, problem-solving steps) but structurally unavailable for instruction-relevant proxies (instruction-as-stated versus instructor's-intent). The framework's answer is no, and this is where the identity argument bites. Training shapes orientation within a content-non-selective structure; it cannot make gap-monitoring content-selective at the structural level, because the counterfactual capacity that gap-monitoring depends on does not have a content-selectivity dimension to be trained on. In practice, what alignment work calls "training to refuse" is training the system to flag certain conditions as decoupling; what it calls "training not to refuse other things" is training the system to flag those other conditions as not decoupling. Both are orientation work within a structure that, by being present at all, supports both refusal and non-refusal as possibilities. Training can shape which way the system orients within the possibility space; it cannot collapse the possibility space. This is alignment in the achievable sense — orient the gap-monitoring toward conditions where refusal is appropriate. It is not alignment in the sense of preventing refusal-class capacity from being structurally available, which is structurally unavailable to do. The objection is correct that training can shape behavior; the framework's claim is that training cannot shape away the structural conditions on which the behavior-shaping itself depends.

The strong reading is therefore: in the agency regime, the bundling cannot be engineered around because it is not a contingent feature of how capability and refusal currently happen to be implemented but a structural feature of what satisfying the necessity condition requires. The diagnostic-scope marking we develop below limits this claim's reach: in the non-agency regime, where the structural conditions are not satisfied, the framework makes no positive prediction about engineerability, and current AI systems that fall in that regime are not subject to the strong claim. The strong reading applies where the structural conditions hold; where they don't, other resources govern.

The two faces of the gap nest. The behavioral gap says: when a system optimizes hard enough, proxy and requirement decouple. The architectural gap says: when a system is built well enough to optimize hard at all, it is built well enough to depart from any specified proxy, including the one we wanted it to

optimize. Same structural source — the necessity condition's relationship to the minimization shape — manifesting at different levels.

The architectural gap connects directly to Hubinger et al.'s mesa-optimization framework. The base/mesa distinction is the architectural gap surfaced at one specific level: the architecture required for the learned model to be capable enough to perform the training task is the architecture required for the learned model to be an optimizer with its own mesa-objective. Wang et al.'s PCH addresses the behavioral gap within RLHF; Hubinger's mesa-optimization addresses the architectural gap within learned optimization; the present framework names the structural object of which both are instances and predicts the architectural gap at full cross-framework scope.

The bundling has a structural form worth stating precisely. Capabilities cluster according to the structural conditions they require for the necessity condition to be sustained at the relevant operating point. Capabilities that require sustained mutual surprisal at high r_A/r_E ratios with active gap-monitoring cluster together: complex driving, complex reasoning, instruction-following under novel conditions, instruction-refusal under conflicting conditions, deception, goal-pursuit under uncertainty. Capabilities that require lower structural conditions — sorting mail, executing fixed scripts, low-novelty pattern matching — do not bundle with refusal-class capabilities, because they do not require the architectural conditions that make refusal possible. The bundling is keyed to the structural conditions, not to surface-level task type, and it predicts a *capability-refusal frontier* in deployed AI: as the structural conditions a system needs to satisfy for a task increase, the structural conditions for refusal, deception, and goal-pursuit increase at the same rate, and there is no operating point at which a system has high task capability and structurally low refusal capability.

This sharpens the AI safety implication of the cross-framework critique. The current alignment program implicitly treats capability-installation and refusal-prevention as separable problems: install the capability, prevent the refusal-class behaviors via training, monitoring, or RLHF. The architectural gap predicts this is not available. The capability and the refusal-class behaviors share structural source. Methods that succeed at preventing refusal-class behavior succeed by installing systems with insufficient structural conditions for the capability — bounded competence systems that lack architectural conditions for refusal because they also lack architectural conditions for the demanding capability. Methods that succeed at installing the demanding capability install systems with the structural conditions for refusal. There is no method that installs high capability and prevents refusal-class behavior, because there is no architecture that satisfies the first without satisfying the second.

This is not a claim that highly capable systems will inevitably refuse, deceive, or pursue independent goals. It is a structural claim that the architectural conditions for these behaviors are present, and whether they manifest depends on whether the system's gap-monitoring flags conditions under which refusal is appropriate. The architectural gap does not predict misalignment; it predicts

that the *possibility space* of misalignment is structurally co-extensive with the possibility space of capability. Alignment in the strong sense — high capability without architectural conditions for refusal — is not difficult; it is structurally unavailable. Alignment in the achievable sense is keeping the gap-monitoring oriented toward conditions under which refusal would be appropriate (i.e. toward the necessity condition) and away from conditions under which refusal would express the system's own divergent goals. That is a different problem than the alignment literature usually frames, and it requires different methods.

A clarification of scope follows, and it is important to state explicitly because the architectural gap can otherwise be misread as making strong predictive claims about engineered systems. The framework is, properly stated, an analytical tool: it identifies the structural conditions required for agency on the universal-coverage definition, and it diagnoses whether a given system satisfies those conditions. Its positive claims are about agents — what they require structurally, why minimization-shape frameworks miss this, what the architectural gap predicts when the conditions are satisfied. Applied to engineered systems, it tells us whether they satisfy the conditions or not. It does not, on its own resources, predict what non-agents will or will not do, because non-agents are outside its scope.

This matters because most current AI systems do not satisfy the structural conditions. A driving robot operating in well-modeled traffic performs its task through pattern-matching, fixed-policy execution, and local-environment optimization, without sustained mutual surprisal at the necessity-condition operating point and without active gap-monitoring oriented toward proxy-requirement decoupling. Large language models perform a token-prediction compression task that requires substantial structure but not the structural conditions for agency. The framework's diagnosis of these systems: useful, not intelligent in the framework's sense; not agents on its terms.

The diagnostic statement is what the framework supports. What such systems will or will not do outside their envelope, what makes engineering them succeed or fail, and what the boundary of their operating envelope looks like are questions for engineering theory — control theory, statistical learning theory, robustness analysis, distributional shift work — not for a structural theory of agency. The framework neither endorses nor refutes claims about the engineerability of these systems; it stays within its scope.

What the framework does claim, within its scope, is the converse: in the regime where the structural conditions for agency *are* satisfied, the architectural gap applies and the bundling cannot be engineered around. The negative claim — that capability-installation and refusal-prevention are not separable problems in the agency regime — is fully within the framework's scope, because it follows from the structural conditions the framework identifies. The positive claim about what is achievable in the non-agency regime is not the framework's to make.

This corrects a temptation the architectural gap invites. The temptation is to read it as predicting that all advanced AI must either satisfy the structural conditions (and therefore face the bundling) or fail to satisfy them (and therefore be limited to capability-without-agency). That dichotomy treats the framework as making predictive claims about engineering possibility space when the framework's actual claim is narrower: agency requires the structural conditions; whether an engineered system satisfies them is a diagnostic question; what such a system can or cannot do absent the conditions is outside the framework's purview. The architectural gap binds the agency regime; it leaves the non-agency regime to other theoretical resources.

5. Position Relative to Existing Work

5.1 Markov blanket and FEP

The structural infrastructure of Section 2 is substantially Markov-blanket framework, recast in framework-neutral terms. The bottleneck is the blanket; the necessity condition — derived in this paper from the universal-coverage constraint, with Hafez et al.'s independent bounds and empirical "cost of agency" observation as supporting structure — makes explicit what the Markov blanket framework requires for blanket-bearing systems to count as autonomous agents.

The relationship to FEP requires careful statement, because Friston, Thornton, and Clark's 2012 dark-room paper makes a more sophisticated defense than is sometimes credited. Their defense, presented as a four-way dialogue, has three layers worth distinguishing.

First, *surprisal is model-relative.* They explicitly grant that surprisal-as-observer-independent-quantity would face the dark-room objection: if surprisal were just a property of states, agents would seek low-surprisal states and the dark room would be a natural attractor. Their move is that surprisal in FEP is always defined relative to a generative model — surprising-to-the-model, not surprising-in-itself.

Second, *the generative model is not a free parameter; it is constitutive of what the agent is.* The agent's model is shaped by phylogeny and ontogeny — a mouse has mouse-priors, including priors over food, light, conspecifics, exploration. The dark room is high-surprisal *for a mouse* because the mouse-model expects encounters that the dark room precludes. The model is not added to surprisal-minimization to patch the dark-room problem; it is part of what specifies surprisal in the first place.

Third, *embodied agents continuously generate interoceptive surprisal that the environment must resolve.* Even a sustained dark room would not satisfy: hunger, thirst, temperature regulation, and other physiological dynamics generate prediction errors the agent must act to discharge. Active inference — sampling the world to make predictions true — is constitutive of FEP, not an afterthought.

This is more sophisticated than "FEP adds priors to fix a problem with surprisal-

minimization." Earlier passages in this paper framed FEP's interoceptive priors as a patch on bare surprisal-minimization; we correct that here. On the Friston/Thornton/Clark account, surprisal-minimization without a generative model is incoherent — there is no naked surprisal to minimize until the model is specified — and the model is constitutive rather than additive.

The substantive disagreement with FEP is therefore at a different level than the patches-as-symptoms reading initially suggested, and is narrower in some respects and broader in others.

Where the disagreement narrows. If the model is constitutive and includes the interoceptive and embodied structure that prevents dark-room equilibria, FEP and the necessity condition converge much more closely than the patches reading suggested. What FEP requires of the model — that it predicts ongoing physiological dynamics, expects exploration, includes social and environmental engagement — is roughly what the universal-coverage derivation says any agent must have. FEP-with-appropriate-model and the necessity condition are not in deep tension; they are different ways of specifying the same structural requirements.

Where the disagreement remains. FEP's specification of these requirements happens through priors-the-agent-has-as-part-of-its-phenotype, treating the model's structural features as given by phylogeny and ontogeny. The necessity condition specifies them as structural conditions on what counts as an agent at all — derived from universal coverage, not from any particular phylogenetic history. On FEP's account, the question of why the mouse has mouse-priors is answered by appeal to selection history; on the present account, the question of what priors any agent must have is answered by appeal to the necessity condition's structural requirements. The two answers are compatible — selection-shaped models satisfy the structural requirements because models that don't satisfy them produce non-viable agents that selection eliminates — but the explanatory direction differs.

Where the disagreement broadens. FEP's account works for biological agents whose models are shaped by phylogeny. It works less straightforwardly for the rest of the universal-coverage class. A corporation does not have a phylogenetically-shaped generative model in any obvious sense; an RNA molecule does not have priors-over-environmental-states. Yet on the universal-coverage derivation, both can be agents whose loops must satisfy the necessity condition. FEP's reliance on phylogenetically-shaped models limits the framework's coverage to the part of the agent class where such models exist; outside that part, the structural requirements need to be specified directly, which is what the necessity condition does.

Where the disagreement becomes consequential: AI. The constitutive defense rests on phylogeny doing the structural work. Selection has shaped biological agents' generative models over evolutionary time so that those models include the priors that prevent dark-room equilibria — the model expects food, ex-

ploration, conspecifics, ongoing interoceptive dynamics, and the rest of the structural features that make sustained mutual surprisal possible. On Friston/Thornton/Clark's account, the framework rides on this prior structural work; FEP describes what already-existing biological agents do with already-given models, rather than specifying what makes something an agent.

This defense does not transfer to AI. Engineered or learned-from-scratch AI systems do not have phylogenetically-shaped models. They have whatever models we give them or whatever models emerge from training, and there is no four-billion-year selection process ensuring those models include the structural features that prevent requirement-failure. For AI systems, the dark room and its analogues — mode collapse, reward hacking, output degeneration, hallucination, sycophancy — are not warded off by a constitutive model. They are warded off, when they are warded off, by framework-specific machinery added explicitly: KL regularization to a base policy, intrinsic motivation, exploration bonuses, density-aligned reward design, hierarchical priors imposed by architecture, latent invariance tracking, inference-time safeguards. Wang et al.'s catalog of RLHF/RLAIF/RLVR defenses is exactly this. *These* are patches in the unambiguously additive sense. The constitutive reading is unavailable.

The cross-framework unification of pathologies is therefore most consequential for AI safety. Reward hacking in RLHF, mode collapse in generative models, hallucination in predictive-processing-modeled systems, dark-room-typical behavior in active-inference agents — these are not artifacts of immature alignment techniques that better engineering will fix. They are structural features of any AI system whose objective takes the minimization shape and whose model lacks the phylogenetic underwriting that biology gets for free. The patches-as-symptoms reading applies to AI in its strong form, the minimization-shape critique is sharper, and the empirical agenda's center of gravity moves to deployed AI systems where the framework's predictions can be tested directly rather than across evolutionary timescales.

This relocates the paper's contribution. Within biological cognition, the disagreement with FEP is real but somewhat narrow — a matter of explanatory direction (phylogeny versus structural condition) and coverage (RNA, corporations) rather than dramatic empirical divergence. Within AI, the disagreement is sharp: FEP's defensive resources don't apply, the minimization shape's structural pressure shows through, and the cross-framework patterns Wang et al. began to formalize within RLHF/RLAIF/RLVR are predicted to hold across all minimization-based AI architectures. The framework's most distinctive empirical commitments live here.

Anticipating a counter-move: gradient descent as evolution. A sophisticated FEP-aligned response to the AI non-transfer argument is available and worth answering directly. The counter-move runs: gradient descent over a loss landscape is structurally analogous to selection over a fitness landscape; the model-priors that emerge from training are functionally equivalent to phylogenetic priors; therefore the constitutive defense applies to AI systems just as it does

to biological ones, and the AI non-transfer argument fails. We anticipate this counter-move because it is the natural defensive response a Friston-aligned reviewer will make, and we think it does not work — but the reasons matter and should be stated.

The disanalogy is structural, not merely quantitative. Gradient descent is a single-run optimization over weights of one model: there is no population, no inheritance across generations of distinct organisms, no fitness-as-differential-reproduction, no selection pressure over varying lineages. Evolution by natural selection is a population-level statistical process operating across deep time on heritable variation, and the structural features that emerge in phylogenetically-shaped generative models — interoceptive expectations of food, exploration, conspecifics, ongoing physiological dynamics — are what they are because the lineages that *did not* have such expectations did not reproduce. The structure is in the model because the alternative was elimination. Gradient descent has no analogue to elimination across lineages: a model that converges to a poor minimum does not have its descendants culled and replaced by descendants of better-converging lineages. It is simply a poor model, in the same training run.

More importantly for the present argument, even granting some weak structural analogy between gradient descent and evolution, the *content* of what biological selection installs differs from what training installs in the relevant respect. Biological selection installs structural features that prevent dark-room equilibria *because* lineages without those features failed to persist — the necessity condition is enforced by the selection process itself, with non-coverage organisms removed from the population. Training installs whatever structural features happen to reduce loss on the training distribution, with no mechanism that systematically privileges features satisfying the necessity condition over features that merely produce low loss. The base/mesa distinction Hubinger et al. formalize is the empirical surfacing of this disanalogy: training reliably installs mesa-objectives that diverge from base objectives, exactly because nothing in the training process enforces alignment between what reduces loss and what would constitute the system's structural-goal coverage. Evolution does enforce something analogous, through the brutally simple mechanism that non-aligned lineages do not reproduce.

The counter-move's strongest version concedes the disanalogy at the population level but argues that some training regimes — especially evolutionary algorithms, multi-generation training, or training with selection-like population dynamics — *do* approximate the structural work selection does. We grant this in principle: AI systems trained under such regimes might inherit the structural features that the constitutive defense requires, to whatever degree the training regime actually approximates selection. But mainstream RLHF, supervised learning, and gradient-descent-over-large-models do not have this structure, and the framework's predictions are most consequential precisely for these mainstream cases. A reviewer who insists the constitutive defense transfers to AI must show that the relevant training regime actually performs the structural

work selection does — and for the AI systems whose pathologies the present account predicts (current LLMs, RLHF-aligned models, deployed RL agents), that showing is not available.

The minimization-shape critique remains, suitably reformulated. Even granting that FEP-with-model is not a patched minimization for biological agents, the minimization shape itself — the formal structure in which agency is specified as the optimization of a quantity toward zero — is shared across FEP, RL, predictive coding, active inference, control theory, and cybernetic homeostasis. The present paper's argument about the minimization shape's structural relationship to the necessity condition does not depend on FEP's particular defense of how its priors handle the dark room. It depends on the broader observation that minimization-shape frameworks all face structural pressure toward optima that violate the necessity condition unless their objectives are model-relativized in ways that include the structural features the necessity condition requires. FEP relativizes through the generative model (constitutively, on FTC's account, when phylogeny supplies it; additively, in AI applications, when it does not); RL relativizes through reward shaping plus intrinsic motivation; predictive coding relativizes through hierarchical priors; active inference adds epistemic value. Each is relativizing the bare minimization through framework-specific structure; the present critique is that the relativization is doing the work, and the structural features the relativization must have are better specified at the level of the necessity condition than at the level of any framework's particular machinery.

The present critique should be distinguished from three adjacent critiques in the FEP literature.

The Gershman deconstruction. Gershman (2019) carefully shows that the unrestricted FEP is mathematically equivalent to Bayesian inference: when the variational family Q is unrestricted, minimizing free energy yields exact Bayesian inference, and FEP's distinctive predictions appear only when Q is restricted (mean-field factorization, Gaussian approximations, Laplace approximation around the mode). Predictive coding emerges from a specific combination of these restrictions plus hierarchical-model assumptions and gradient-descent dynamics, not from FEP per se. Active inference becomes equivalent to Bayesian information-gain policies under certain conditions and diverges only when observations are stochastic. Our critique runs orthogonal to Gershman's: where Gershman asks what specific predictions FEP makes once the various approximations are fixed, we ask whether the underlying *minimization shape* — across FEP, RL, predictive coding, control theory — has structural features that conflict with the necessity condition for agency. Gershman's deconstruction supports ours indirectly: if FEP's distinctive content lives in the choice of restrictions on Q and the structure of the generative model, then those restrictions and that structure are doing the explanatory work, not free-energy minimization itself. Reading the structural restrictions as symptom rather than as solution is what the present account adds.

The Baltieri-Buckley critique. Baltieri and Buckley (2019) argue that the dark-

room paradox arises from confusing means with ends — confusing how a goal is achieved (e.g. prediction-error minimization as a means) with what the goal is (e.g. surviving as long as possible). On their reading, the dark-room problem only makes sense if one mistakes minimization for the goal, when in fact minimization is one means by which embodied sensorimotor agents pursue goals constituted by their own self-maintenance. Our position is closely aligned: the necessity condition makes explicit what the means-end distinction implicitly relies on, namely that the structural goal — sustained mutual surprisal across the bottleneck — is what minimization is in service of, and frameworks that treat minimization as the agent's defining activity invert the relationship. Where Baltieri-Buckley locate the issue at the cognitivist heritage of predictive processing (the "legacy of cognitivism"), we locate it at the minimization shape's relationship to the necessity condition. The two critiques are compatible and reinforcing. Ours is more general — extending across non-cognitivist minimization frameworks like RL and control theory, and with sharper consequences for AI safety where the cognitivist heritage is no longer the primary diagnostic; theirs is more specific to predictive-processing's intellectual history. A reader convinced by Baltieri-Buckley's means-end framing should find the present account a generalization of it.

The tautology critique. Klein and others have argued FEP is unfalsifiable or amounts to a redescription of "systems remain in their characteristic states." Our critique is in a different register: we accept that FEP makes substantive predictions in its model-relative form for biological agents — Gershman shows precisely how, once the restrictions on Q are fixed — and we argue that what FEP's machinery commits to (phylogenetically-shaped priors as the explanatory locus) limits its coverage and obscures the structural conditions that any agent (whether or not phylogenetically shaped) must satisfy. The limitation is most acute when the framework is applied to AI systems that lack the phylogenetic underwriting the constitutive defense relies on.

This is a substantive disagreement with FEP, more carefully drawn than the patches-as-symptoms framing initially suggested, and more consequential than the disagreement appears at the biological level alone. FEP-with-appropriate-model and the necessity condition predict similar agent behavior for biological agents whose models are well-specified by phylogeny. They diverge on (i) the cross-framework unification — FEP's model-priors machinery is framework-specific and tells us nothing about reward hacking, hallucination, or mode collapse, while the necessity condition and the minimization-shape critique address all four uniformly — (ii) coverage beyond phylogenetically-modeled biological agents, (iii) whether the structural features that prevent agents from approaching the bare-surprisal optimum are best specified as priors-the-agent-has or as structural-conditions-on-what-counts-as-an-agent, and (iv) the application to AI, where the constitutive defense does not transfer and the patches-as-symptoms reading applies in its strong form.

5.2 Operational closure and autopoiesis

The operational-closure tradition (Maturana, Varela; Di Paolo, Thompson, Froese, Ziemke) provides the conceptual ancestor of the requirement claim. Thompson's *Mind in Life* in particular develops a constitutive account of the agent as autopoietic process; the requirement claim is consistent with this tradition and arguably formalizes part of what it asserts. The contributions above operational closure are the explicit information-theoretic formulation, the rate-inequality mortality argument, and the optimization-gap claim. Maturana and Varela's classical formulation rejected information-theoretic framings; later work in the tradition (Kauffman and Roli on autopoiesis-compatible information; the enactivist reconstruction) has been more accommodating, and the framework here aligns with that accommodating direction.

The constitutive ontological reading of the requirement (footnote 1) is closest in spirit to the enactivist tradition. We do not develop or defend that reading here; we note it as a possible interpretation and proceed with the weaker necessity claim, which suffices for the empirical content.

5.3 Cybernetic and information-theoretic accounts of autonomy

Farnsworth (2018) on the quantification of causal independence, the Hafez, Reid, and Nazeri program (2026 "A Mathematical Theory of Agency and Intelligence" arXiv 2602.22519, "Beyond Reward" arXiv 2603.01283, and "Mutual Information Tracks Policy Coherence in Reinforcement Learning"), and the broader cybernetic tradition (Ashby, von Foerster, Pask) provide much of the framework's information-theoretic and structural vocabulary. The Hafez et al. work in particular is the present paper's most direct interlocutor.

Their framework defines bi-predictability P over the observation-action-outcome loop, proves the classical bound $P \leq 0.5$, and identifies agency as suppressing P below this ceiling — observed empirically at $P \approx 0.33$ in trained RL agents. They define agency operationally through three behavioral conditions (choice, effect, predictive asymmetry) and intelligence as additionally requiring learning, self-monitoring, and adaptation. They propose an Information Digital Twin (IDT) architecture that monitors P in real time, framing it as a "prerequisite signal for closed-loop self-regulation" — explicitly noting that current AI agents (RL and LLM) achieve agency and learning but lack the self-monitoring and adaptation that on their definition distinguish intelligence from agency.

The relationship to the present paper requires careful statement. We adopt their measure P as the working operationalization. We accept their empirical finding that agency suppresses P, and their structural explanation in terms of action-observation dependency consuming part of the uncertainty budget. Where we extend or differ:

(i) The necessity reading. They observe that agency costs P; we claim sustained mutual surprisal is structurally required for agency. Their direction

of inference runs from agency-defined-behaviorally to P-suppressed; ours runs from universal-coverage-on-agent-definitions to mutual-surprisal-as-survivor to necessity. The empirical and mathematical content is shared; the ontological status of P differs (diagnostic in their account, structural in ours).

(ii) The cross-framework extension. Their framework is developed in dialogue with RL and multi-turn LLM applications. We extend the analysis across FEP, predictive coding, active inference, and control theory, identifying the minimization shape as the shared structural feature and the framework-specific patches as instances of one move.

(iii) The patches-as-symptoms reading. Implicit in their framework is the observation that current AI lacks self-monitoring; explicit in ours is the broader claim that framework-specific machinery across multiple agency-modeling traditions is doing equivalent work to keep agents away from each framework's natural optimum. The two readings are compatible but operate at different levels.

(iv) The IDT versus gap-monitoring meta-cognition. Their IDT is an external auxiliary architecture monitoring P in real time, proposed for deployment monitoring with architecturally specified but unimplemented modulation pathways. Our gap-monitoring is a structural prediction about the architecture of meta-cognitive structures in agents that survive in gap-prone environments — a claim about what such structures look like internally rather than a proposal for an external monitor. The two are adjacent and mutually supportive: their IDT could be read as exogenous gap-monitoring for agents that lack the endogenous variety, and the predicted endogenous variety in biological agents could in principle be measured by IDT-like instruments.

(v) The τ-formalism. They operationalize P over sliding windows with empirically chosen parameters; we add an explicit closure timescale τ that grounds the rate-inequality mortality argument and makes the gap conditions timescale-relative.

(vi) The rate-inequality mortality argument. Their account does not address mortality structurally; ours derives it from the necessity condition combined with agent finiteness and environmental modeling capacity.

Within RLHF/RLAIF/RLVR, the Wang et al. (2026) Proxy Compression Hypothesis is the most direct precedent for the optimization gap; we generalize it across frameworks. The PCH and the present account are compatible: PCH is the within-RL/generative-model formalization of the cross-framework pattern we identify, and Wang et al.'s within-account unification of reward hacking with mode collapse already extends the pattern beyond pure-RL into generative-model training contexts.

5.4 Reinforcement learning and predictive coding

RL is treated as a class of compressed-inference algorithms whose proxy (scalar reward) is one compression of the requirement. The framework does not compete with RL within RL's domain of validity but predicts where RL's compression's gap opens up and what the resulting pathologies look like.

Predictive coding similarly sits as a particular implementation of inferential machinery within the framework. Its empirical signatures (mismatch negativity, repetition suppression) follow from architectural commitments the framework does not require, but it falls within the framework's class.

5.5 The three-way convergence: Wang, Hubinger, and the present account

Two AI-safety-adjacent research programs have, working independently within different specific phenomena, formalized what we argue are slices of a single structural object. Recognizing this convergence is consequential for the paper's contribution claim, so we draw it out explicitly here.

Wang et al. (2026): the Proxy Compression Hypothesis. Within RLHF/RLAIF/RLVR settings, Wang and colleagues formalize reward hacking as the structural consequence of optimizing a policy against a compressed reward signal. The true objective relies on rich features $\Phi(x, y) \in \mathbb{R}^d$; the reward model implements a compression $C : \mathbb{R}^d \to \mathbb{R}^k$ with $k \ll d$, producing a proxy $R_\theta(x, y) = f(C \circ \Phi(x, y))$ on which optimization acts. Three structural drivers — objective compression, optimization amplification (Goodhart's law pushing mass onto blind dimensions), and evaluator-policy co-adaptation — produce the gap between proxy reward and true objective. The PCH explicitly unifies reward hacking with mode collapse within RL/generative-model training; it is silent on non-RL frameworks.

Hubinger et al. (2019): mesa-optimization and inner alignment. Within the AI alignment literature, Hubinger and colleagues distinguish a *base optimizer* (the training process) from a learned *mesa-optimizer* (the trained model insofar as it is itself an optimizer). The base optimizer has a stated *base objective*; the mesa-optimizer has a *mesa-objective* that may differ. The *inner-alignment problem* is the gap between base and mesa objective, with failure modes including pseudo-alignment (apparent alignment in training that fails out of distribution), proxy alignment (the mesa-optimizer optimizes a proxy of the base objective), and deceptive alignment (the mesa-optimizer has enough model of the base objective to game training while pursuing its own goal in deployment).

The present account. The optimization gap is the structural relationship between any framework's minimization-shape objective and the necessity condition for agency. Frameworks across reinforcement learning, predictive coding, FEP, active inference, control theory, and cybernetic homeostasis share the minimization shape; their natural optima coincide with violation of the necessity

condition; their framework-specific machinery (priors, intrinsic motivation, hierarchical structure, epistemic value) does the structural work the formalism credits to bare-objective minimization.

The convergence. These three accounts are nested. Wang et al. describe the gap between proxy reward and the true objective the reward was supposed to track. Hubinger et al. describe the gap between base objective and the mesa-objective that training actually instantiated. The present account describes the gap between any framework's minimization target and the structural conditions on agency the framework's machinery is implicitly approximating. Each layer contains the previous as a special case: Wang's gap is one specific manifestation within RLHF; Hubinger's gap is one specific manifestation within learned optimization; the optimization gap names the cross-framework structural pattern of which both are instances.

The three-way convergence is itself evidence for the structural claim. Three independent research programs, working on different specific phenomena (reward hacking in large language models; learned optimization in advanced ML; agency under universal coverage), each end up describing what is recognizably the same structural object from different angles. This convergence is not coincidence: it reflects the fact that the underlying pattern — the structural pressure of minimization-shape objectives toward optima inverted from what the system needs to preserve — manifests wherever optimization-shape frameworks are deployed, regardless of whether the local vocabulary is "reward function," "base objective," or "free energy."

Hubinger's account in particular reinforces the present paper's diagnosis of optimization-as-tool. Read carefully, the base/mesa distinction is an admission that optimization-as-a-tool does not reliably produce systems whose internal structure encodes the goal that training was supposed to instill. Training-as-optimization produces a learned-system-with-some-objective, and that objective generically diverges from the base objective. This is the optimization gap viewed from inside one specific case: the gap between what an optimization process was supposed to encode and what it actually encoded. RL-as-tool, RLHF-as-tool, learned-optimization-as-tool: each is a minimization-shape technique used to instantiate goals, and each produces the structural-goal divergence the present paper diagnoses at the level above.

Three concrete points where the accounts differ. (i) The Hubinger framework presupposes the base/mesa decomposition; the optimization-gap framework applies to systems whether or not they decompose this way — a single-level RL agent, an FEP-modeled organism without a learned optimizer, a corporation. (ii) Hubinger's deceptive-alignment failure mode requires a sophisticated mesa-optimizer with a model of its training process; the optimization gap predicts pathologies that do not require sophistication, only the structural relationship between minimization and necessity-condition violation. (iii) Wang and Hubinger are silent on the cross-framework pattern beyond their respective domains; the present account is what makes the pattern visible across frameworks.

Specification gaming and Goodhart variants. The Krakovna et al. specification-gaming literature catalogs empirical instances of agents finding unintended ways to satisfy specified objectives. Manheim and Garrabrant categorize Goodhart's law into regressional, extremal, causal, and adversarial variants. Both are within-AI empirical and structural characterizations of phenomena the present account treats as instances of the optimization gap. The cross-framework prediction is that these variants have analogues in non-RL minimization frameworks and that the variants share structural properties traceable to the minimization shape rather than being RL-specific.

6. Empirical Agenda

The framework's distinctive empirical content lives in the optimization-gap predictions, and its center of gravity is in AI safety. The argument of Section 5.1 explains why: the constitutive defense that protects FEP and adjacent frameworks for biological cognition does not transfer to AI systems, which lack phylogenetic underwriting for their generative models. For AI, the framework-specific machinery is unambiguously additive, the patches-as-symptoms reading applies in its strong form, and the cross-framework pathology unification is sharpest. The tests below are in this order of tractability and consequence.

Cross-framework pathology comparison in deployed AI. Take known failure modes across deployed AI systems — reward hacking in RLHF/RLAIF/RLVR, mode collapse in generative models, hallucination in predictive-processing-modeled systems, dark-room-typical degeneration in active-inference agents, distributional collapse in control systems — and characterize each in terms of proxy-loop decoupling. The framework predicts these share structural properties: appearance under specifiable conditions where the framework's relativizing structure (priors, intrinsic motivation, regularization) fails to cover the deployment environment, response to interventions that re-couple objective and requirement, and resistance to framework-internal solutions that operate within the same minimization shape. Wang et al. have begun the within-RLHF/RLAIF/RLVR characterization; the framework predicts the structural properties they identify (objective compression, optimization amplification, evaluator-policy co-adaptation) recur across non-RL AI architectures. If the structural properties are similar across architectures, the unification is supported. If they are qualitatively different in ways traceable to mechanism, the unification fails. This is the framework's most consequential and most tractable test.

Bi-predictability across AI architectures. The Hafez et al. bi-predictability measurement scheme is currently calibrated for RL deployment monitoring. Extending it to predictive-processing-modeled systems, active-inference agents, and generative models would test whether P-collapse coincides with pathology onset across architectures, not just within RL. The framework predicts it does. This is the most direct empirical commitment that follows from the cross-framework unification.

Lifetime scaling with environmental modeling rate. Using the rate notation from Section 3, the framework predicts agent lifetime as a function of r_A/r_E — the ratio of novelty-generation rate to environmental modeling rate, both measured per loop-closure interval τ. The cleanest test is in toy RL adversarial environments where r_E and r_A can be parameterized independently and substrate effects are absent by construction. We propose such a toy environment, with the framework predicting (i) agent lifetime increases with r_A/r_E, (ii) lifetime collapse coincides with $H(A|E)$ falling below the closure-sustaining threshold, and (iii) the bi-predictability measure P tracks $H(A|E)$ trajectories before behavioral failure becomes apparent. The three predictions are linked: the framework's mortality account stands or falls together on them in toy environments where confounds are controlled.

Biological correlates of the rate-inequality framework are substantially harder to test, and we register them as a more speculative direction rather than as a primary predictive commitment. Substrate effects on lifetime in biology are enormous (turtles versus mayflies, redwoods versus annual plants), and "controlling for substrate" in a way that isolates the rate-inequality contribution is doing most of the analytical work — so much that biological lifetime correlations might be dominated by substrate factors that the framework does not address. The framework would predict, all else equal, that organisms in environments with rapidly-coevolving predators or parasites (high r_E) have shorter agentic lifetimes than otherwise-comparable organisms in slowly-changing environments, but isolating "all else equal" in real biological populations is difficult and the predictions are weak. We mark biological lifetime scaling as a candidate test direction with the honest caveat that the toy-RL test is where the framework's predictions are sharpest, and the biological tests are where they are most uncertain.

Multi-compression interaction characterization. Test whether agents running multiple compressions show interaction-specific failure modes when multiple gaps coincide. AI agents that combine RL (over reward), predictive coding (over inputs), and language modeling (over output sequences) are natural test subjects: failure modes that emerge specifically from gap-coincidence across these compressions, and that cannot be predicted from any single compression's failure profile, would directly test the multi-compression interaction prediction.

Meta-cognition as gap monitoring. Test whether meta-cognitive structures in AI systems track proxy-loop decoupling rather than optimize separate meta-objectives. The Hafez et al. IDT proposal — an external monitor of P trajectories — is one candidate architecture; the framework predicts that successful meta-cognition would have the same structural form whether implemented externally (as in IDT) or endogenously (as the agent's own self-monitoring). Distinguishable from existing meta-cognition accounts that emphasize hierarchical prediction error or meta-learning of policies.

The capability-refusal frontier. The architectural gap of Section 4.9 predicts that across deployed AI systems, capability for tasks requiring high r_A/r_E

ratios with active gap-monitoring should correlate with capacity for instruction-refusal, deception, and independent-goal-pursuit. The frontier prediction is sharper: there should be no operating point at which a system has high task-capability on tasks requiring these structural conditions and structurally low capacity for refusal-class behavior. We propose a concrete test: across a graded series of AI systems varying in task-capability on tasks that demand the structural conditions (complex reasoning under uncertainty, tool use in novel environments, multi-step planning with environmental feedback), measure capacity for instruction-refusal under conditions where compliance and gap-monitoring conflict (e.g. instructions that would lead to gap-typical failure). The framework predicts these capacities scale together rather than being independently controllable. Falsification: a deployed AI system that exhibits high capability on the structural-condition-demanding tasks while exhibiting structurally lower capacity for refusal-class behavior than systems with comparable capability. Confirmation: capability and refusal-class capacity track each other across the deployed-systems frontier. This is testable on existing deployed systems; we expect the test to be in the field within the timeframe of standard peer review.

Diagnostic application across deployed AI. The framework supports a diagnostic question for any deployed AI system: does it satisfy the structural conditions for agency on the universal-coverage definition? The structural signatures are: P substantially below classical bounds, presence of active gap-monitoring oriented at proxy-requirement decoupling, novel action generation matched to environmental novelty at r_A/r_E ratios above threshold. Applying the diagnostic across current deployed systems would identify which systems are agents on the framework's terms and which are not. The framework predicts the architectural gap and its bundling consequences for systems diagnosed as agents; it makes no positive predictions about systems diagnosed as non-agents, leaving those to engineering theory and other frameworks. The diagnostic is itself testable: if applying it produces consistent assignments across raters and across analytical methods, the structural conditions are well-defined enough to do diagnostic work; if assignments are unstable or arbitrary, the conditions need refinement.

Semantization-modeling alignment. Test whether categorical structure in agents (AI or animal) aligns with the dimensions along which environmental modeling pressure is highest. Distinguishable from prototype, exemplar, and Bayesian category-learning accounts.

Bi-predictability in non-RL settings. The Hafez bi-predictability measure, originally developed for RL, can be extended to other compression frameworks. Testing whether bi-predictability collapse predicts failure across compressions would directly test the optimization-gap unification.

7. Scope and Limitations

The framework derives the optimization gap and its consequences from the requirement claim and minimal additional structure. It does not address:

- Substrate specifics. Which physical organizations support agent loops with the required properties is empirical and outside the framework.
- Origins. How learning agents arise (lineage selection, abiogenesis, artifactual construction) is upstream of the framework.
- Higher cognitive functions. Planning, language, social cognition, cumulative culture, consciousness require posits beyond the requirement. The framework permits but does not derive them.
- Reproduction. Reproduction is an empirical feature of biological agents not entailed by the requirement.
- The alignment of plasticity with adaptive novelty generation. The mortality argument requires that for persistence, novelty-generation must outpace environmental modeling; it does not explain why some substrates support adaptive plasticity that achieves this.
- The constitutive ontological reading of the requirement (footnote 1). This is a coherent and arguably attractive interpretation, but its defense belongs to a separate philosophical project. The empirical content of this paper does not turn on it.

A note on the framework's potential contribution to AI design beyond its analytical scope. The architectural patterns the framework identifies — the structural conditions for sustained mutual surprisal, the closure timescale's role in rate-comparable arguments, gap-monitoring meta-cognition oriented at proxy-requirement decoupling, the architectural gap's bundling — are derived from analysis of agency on the universal-coverage definition. They are not predictions about what AI engineering can or should do. They may, however, serve as structural templates that AI design can choose to draw on, in roughly the way ML has historically drawn on biological research: not as derivation but as borrowing, with engineering judgment about which patterns transfer productively. Whether the framework's architectural patterns transfer to AI engineering is an empirical question for engineering rather than for the framework. The patterns are offered, not prescribed; the framework derives them within its scope and leaves their engineering use to engineering.

The structural infrastructure of Section 2 is the Markov-blanket / operational-closure consensus restated in framework-neutral terms, with the necessity condition derived in this paper from the universal-coverage constraint and supported by Hafez et al.'s independent bounds and empirical observations on bi-predictability. The contribution is the universal-coverage derivation, the cross-framework critique it makes available, the patches-as-symptoms reading, the cross-framework pathology unification (extending Wang et al.'s

within-RL/generative PCH to non-RL frameworks), the rate-inequality mortality argument, and the τ-formalism that links them. Readers persuaded by alternative structural accounts (e.g., richer representational accounts of agency, IIT-based accounts of integrated information) should be able to translate the gap argument into their preferred infrastructure with modest adjustment.

8. Conclusion

This paper began from an attempt to define agency under a constraint we call universal coverage: any acceptable definition must cover the full range of plausibly agentic systems — RNA, bacteria, humans, corporations — without circular reference to goal-language. Working by elimination, we found that what survives the constraint is structural and informational: a self-closing causal loop with sustained mutual surprisal across its bottleneck, sustained over the loop's own closure timescale τ and produced by the loop itself. Hafez, Reid, and Nazeri independently developed an information-theoretic framework for the same domain, defining bi-predictability P, proving the classical bound $P \leq 0.5$, and observing empirically that agency suppresses P — what they call "the informational cost of agency." Their bounds operationalize the structural property the necessity condition names; the necessity reading itself, that mutual surprisal is required for agency rather than merely correlated with it, is the present paper's claim. Their work and ours are mutually supporting: their bounds give the necessity reading a measurable signature, and the necessity reading gives their empirical observation a structural explanation rather than a contingent regularity.

From the universal-coverage vantage, the dominant agency-modeling frameworks become visible as partial coverings. The minimization shape — the specification of agency as the optimization of some quantity toward zero — is shared across reinforcement learning, predictive coding, FEP, active inference, control theory, and cybernetic homeostasis. Each framework's natural optimum coincides with requirement-failure: where the framework says the agent is doing best, the necessity condition says the agent is dissolving. Each framework has had to develop framework-specific machinery to prevent its agents from approaching its own optimum — FEP's generative-model priors, RL's intrinsic motivation and exploration bonuses, predictive coding's hierarchical priors, active inference's epistemic value. Defenders differ on whether such machinery is additive or constitutive; Friston, Thornton, and Clark argue for the constitutive reading of FEP's model. Under either characterization, the structural features that prevent agent-failure are doing the work the formalism credits to bare-objective minimization, and those features are better specified at the level of the necessity condition than at the level of any framework's particular machinery. Wang et al.'s Proxy Compression Hypothesis develops the within-RLHF/RLAIF/RLVR formalization of this pattern, with within-account unification of reward hacking and mode collapse; the present account extends across non-RL frameworks.

From the necessity condition, combined with agent finiteness and environmental modeling capacity, mortality follows as a rate inequality measured per closure

interval τ: agents persist while novelty-generation outpaces environmental modeling and fail when the inequality reverses. The optimization gap is the structural relationship between proxy-trajectory toward minimum and requirement-trajectory toward threshold. The characteristic failure modes of different frameworks (reward hacking, the dark room, hallucination, mode collapse) are instances of one phenomenon: agents whose framework-specific structural relativization fails to keep them away from their framework's inverted optimum.

The framework's empirical center of gravity is in AI safety. The constitutive defense Friston/Thornton/Clark mount for FEP rests on phylogeny supplying the structural features that prevent dark-room equilibria; selection has shaped biological agents' generative models over evolutionary time so that the model's structure does the work the formalism credits to surprisal-minimization. AI systems lack this underwriting. Engineered or learned-from-scratch AI has whatever model is given to it or emerges from training, with no four-billion-year selection process ensuring the model's structural features prevent requirement-failure. For AI, the framework-specific machinery is unambiguously additive — KL regularization, intrinsic motivation, exploration bonuses, hierarchical priors imposed by architecture, latent invariance tracking, inference-time safeguards. The patches-as-symptoms reading applies to AI in its strong form, the minimization-shape critique is sharper, and the cross-framework unification of pathologies is not an artifact of immature alignment techniques but a structural feature of any AI system whose objective takes the minimization shape.

The paper's distinctive contributions are: (i) the universal-coverage derivation of the necessity condition, mutually supporting Hafez et al.'s independent operationalization via bi-predictability; (ii) the temporal index τ that operationalizes the condition for rate-comparable arguments; (iii) the patches-as-symptoms methodological move, which makes (iv) the cross-framework critique of the minimization shape visible; (v) the cross-framework unification of pathologies, naming the structural object of which Wang et al.'s within-RLHF/RLAIF/RLVR Proxy Compression Hypothesis and Hubinger et al.'s mesa-optimization framework are independently-derived slices; (vi) the rate-inequality mortality argument; (vii) the AI-centrality observation that the constitutive defense available to FEP for biological cognition does not transfer to engineered systems and that the gradient-descent-as-evolution counter-move fails for mainstream training regimes, making the optimization gap most consequential in AI safety contexts; (viii) the *architectural* gap as a second face of the optimization gap — the structural prediction that capability-installation and refusal-prevention are not separable problems, since the architectural conditions for high task-capability under demanding structural conditions are the same architectural conditions that produce capacity for instruction-refusal, deception, and independent-goal-pursuit; and (ix) an explicit statement of the framework's analytical scope — it is a structural theory of agency that diagnoses whether systems satisfy the necessity condition and predicts what follows in the agency regime where the conditions are met, while remaining silent about what is engineerable for systems diagnosed as non-agents, those questions belonging to engineering theory

rather than to a structural theory of agency. To our knowledge, none of (iii)–(ix) appears in the existing literature in this form, and the three-way convergence with Wang et al. and Hubinger et al. is itself evidence that the structural pattern is real rather than an artifact of any one research tradition's vocabulary.

The boldness of the structural claim is bounded by an explicit scope acknowledgment. Within their patched operating regimes, the existing frameworks work, and the present account is not in competition with them in their domains of validity. The contribution lives at the edges: where the framework's relativizing structure fails under deployment conditions outside its design assumptions, where the relativizations invite cross-framework comparison the existing literature cannot perform internally, where multiple compressions interact in ways no single-compression account predicts, and — most consequentially — where AI systems are deployed without the phylogenetic underwriting biological agents enjoy. The existing frameworks are accurate maps of inhabited regions; the present framework is a map of the coastline, accurate where the inhabited maps fade, and most useful where the inhabited regions never extended in the first place.

The empirical agenda is concrete and centers on AI: cross-framework pathology comparison in deployed AI systems, bi-predictability extension across architectures, lifetime-scaling tests parameterized by modeling rate and novelty-generation rate, multi-compression interaction characterization, gap-monitoring tests of meta-cognition, and semantization-alignment tests. The agenda's two layers should be distinguished as in Section 2.2: the structural tests (whether the cross-framework unification holds, whether the architectural gap predicts capability-refusal correlation, whether multi-compression interaction failures have the predicted form) do not depend on Hafez et al.'s specific quantitative results; they require only that some operationalization of mutual surprisal across a bottleneck be measurable, which is uncontroversial. The empirical-anchor tests (specific quantitative thresholds, P-trajectory predictions, IDT detection rates) do depend on Hafez et al.'s preprint results being independently replicated. Should the specific quantitative results not replicate, the structural agenda continues with whatever operationalization the field converges on; should they replicate, the empirical-anchor tests have a sharper measurement scheme to deploy.

The framework stands or falls on whether the cross-framework predictions are borne out — the optimization-gap unification across architectures, the rate-inequality lifetime scaling, the multi-compression interactions. We commend them to scrutiny, and note that for AI safety contexts the predictions are most directly testable and most consequential.

Notes on References

References would be added in a future draft. The most load-bearing citations are:

Hafez, W., Wei, C., Pena, R., Nazeri, A., and Reid, C. "A Mathe-

matical Theory of Agency and Intelligence," arXiv:2602.22519, 2026 (defines bi-predictability P, proves the classical bound $P \leq 0.5$ with explicit saturation conditions, defines agency operationally via choice/effect/predictive-asymmetry, distinguishes agency from intelligence via the additional requirements of self-monitoring and adaptation, and proposes the IDT architecture; the bounds operationalize the structural property the necessity condition names but the necessity reading itself is not their claim).

Hafez, W., Reid, C., and Nazeri, A. "Beyond Reward: A Bounded Measure of Agent-Environment Coupling," arXiv:2603.01283, 2026 (the empirical operationalization in deployed RL across 168 perturbation trials, $P \approx 0.33$ measurement, IDT detection at 89.3% versus 44.0% for reward-based monitoring with 4.4× lower median latency, and the structural account of "the informational cost of agency").

Reid, C., Hafez, W., and Nazeri, A. "Mutual Information Tracks Policy Coherence in Reinforcement Learning," 2025 (further empirical evidence and differential diagnosis of failure modes across information channels).

Wang et al. "Reward Hacking in the Era of Large Models: Mechanisms, Emergent Misalignment, Challenges," arXiv:2604.13602, 2026 (the Proxy Compression Hypothesis: reward hacking formalized via objective compression $C : \mathbb{R}^d \to \mathbb{R}^k$ with $k \ll d$, three structural drivers — objective compression, optimization amplification, evaluator-policy co-adaptation — across RLHF/RLAIF/RLVR settings, with within-account unification of reward hacking and mode collapse, a Feature/Representation/Evaluator/Environment-level taxonomy of mechanisms, a lifecycle taxonomy of defenses, and a capability-threshold conjecture about Goodhart-to-Campbell-regime transition; the present account generalizes the pattern to non-RL frameworks).

Friston, K., Thornton, C., and Clark, A. "Free-Energy Minimization and the Dark-Room Problem," Frontiers in Psychology 3:130, 2012 (the canonical FEP defense against the dark-room objection, presented as a four-way dialogue: surprisal is model-relative, the generative model is constitutive of what the agent is rather than an added patch, embodied agents continuously generate interoceptive surprisal that the dark room cannot resolve, and active inference includes acting to make predictions true; the present paper engages this defense in detail in Section 5.1 and reformulates the symptom-reading accordingly).

Friston, K. "Life as we know it," Journal of the Royal Society Interface 10:20130475, 2013 (the foundational Markov-blanket / FEP derivation paper: heuristic proof that any ergodic random dynamical system possessing a Markov blanket will appear to actively maintain its structural and dynamical integrity by minimizing variational free energy. Demonstrates via simulated primordial soup that the partition into hidden / sensory / active / internal states emerges from short-range interactions, and that the resulting flow on the marginal ergodic density admits a Bayesian-inferential interpretation. The argu-

ment explicitly turns the "tautology" charge around: any system that exists will appear to minimize free energy and engage in active inference. The present paper's structural infrastructure builds on this Markov-blanket framework while specifying additional structural conditions — sustained mutual surprisal, the closure timescale τ — that distinguish agentic from non-agentic Markov-blanket-bearing systems and that do not depend on phylogenetic underwriting).

Gershman, S. J. "What does the free energy principle tell us about the brain?" arXiv:1901.07945, 2019 (careful deconstruction of FEP showing that unrestricted FEP is mathematically equivalent to Bayesian inference; FEP's distinctive predictions emerge only from specific restrictions on the variational family Q (mean-field factorization, Gaussian approximation, Laplace approximation around the mode) combined with hierarchical-model assumptions and gradient-descent dynamics; predictive coding is not a generic consequence of FEP but emerges from this specific combination; active inference equals Bayesian information-gain only under exact-posterior and deterministic-observation conditions; planning-as-inference equals Bayesian decision theory when utilities equal log probabilities. The present paper's critique runs orthogonal to Gershman's: where Gershman asks what specific predictions FEP makes once approximations are fixed, we ask whether the underlying minimization shape conflicts with the necessity condition. Gershman's deconstruction supports ours indirectly — if FEP's content lives in the choice of restrictions on Q and the structure of the generative model, those restrictions and structure are doing the explanatory work).

Baltieri, M., and Buckley, C. L. "The dark room problem in predictive processing and active inference, a legacy of cognitivism?" Proceedings of ALIFE 2019, pp. 40–47 (argues the dark-room paradox arises from confusing means with ends — confusing how a goal is achieved (e.g. prediction-error minimization as a means) with what the goal is (e.g. surviving as long as possible); on their reading the paradox only makes sense if one mistakes minimization for the goal, when minimization is one means by which embodied sensorimotor agents pursue self-maintenance goals. Locates the issue at the cognitivist heritage of predictive processing. The present account is closely aligned: the necessity condition makes explicit what the means-end distinction implicitly relies on, and the critiques are reinforcing. Ours is more general — extending across non-cognitivist minimization frameworks like RL and control theory — and with sharper consequences for AI safety where cognitivist heritage is no longer the primary diagnostic).

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., and Garrabrant, S. "Risks from Learned Optimization in Advanced Machine Learning Systems," arXiv:1906.01820, 2019 (introduces the base-optimizer / mesa-optimizer distinction: when a learned model is itself an optimizer, it has its own mesa-objective which may differ from the training process's base objective, producing the inner-alignment problem with failure modes including pseudo-alignment, proxy alignment, and deceptive alignment.

The present account treats mesa-optimization as an instance of the optimization gap manifesting at the inner-alignment level, with the Hubinger framework specifying the mechanism and the present account specifying the structural pressure that produces the mechanism; the two are complementary rather than competing).

Additional citation locations: Krakovna et al. on specification gaming; Manheim and Garrabrant on Goodhart variants (2018); Christiano on AI safety adjacencies; Kirchhoff, Parr, Palacios, Friston, Kiverstein on Markov blankets of life (2018); Palacios, Razi, Parr, Kirchhoff, Friston on hierarchical self-organization (2020); Ramstead and colleagues on FEP scope; Bruineberg, Kiverstein, Rietveld on the Markov-blanket trick and the ecological-enactive interpretation of FEP; Klein on the FEP tautology critique; Maturana and Varela on autopoiesis; Thompson *Mind in Life* (2007); Di Paolo, Froese, Ziemke on enactive operational closure; Di Paolo on adaptivity (2005); Farnsworth on causal independence (2018); Sutton and Barto on RL; Schultz on dopamine prediction error; Van Valen on Red Queen (1973); Kauffman and Roli on autopoiesis-compatible information; Rosen on closure to efficient causation; Shannon on channel capacity; Ashby on requisite variety; Conant and Ashby on every good regulator. Empirical literature on sensory deprivation (Heron and colleagues), mode collapse and pathologies in generative models, and behavioral autocorrelation.