Autonomy Suppression in Hierarchical Multi-Agent Systems: A Unifying Systems-Theoretic Framework

Boris Kriger[1,2]

[1]*Information Physics Institute, Gosport, Hampshire, United Kingdom*
boris.kriger@informationphysicsinstitute.net

[2]*Institute of Integrative and Interdisciplinary Research, Toronto, Canada*
boriskriger@interdisciplinary-institute.org

Abstract

The problems of distributional shift in offline reinforcement learning, exploration suppression under safety constraints, reward hacking under misaligned incentives, and fragility under volatility deprivation have each been studied extensively in their respective literatures. This paper argues that these apparently distinct pathologies are consequences of a single structural parameter: the *intervention strength* α of a supervisory agent in a hierarchical system. We develop a formal framework—grounded in constrained Markov decision processes, information theory, and distributional robustness—that provides a unified vocabulary for analysing this phenomenon across domains. The framework yields a parametric family of intervention operators, a formal characterisation of the critical threshold beyond which protection becomes counterproductive, and an architecture (*envelope governance*) that replaces trajectory control with boundary constraint. Our contribution is primarily one of synthesis and formalisation: we show that developmental psychology's zone of proximal development, curriculum learning's difficulty scheduling, autonomous driving's safe-envelope control, and organisational scaffolding are instantiations of a single abstract scheme. The resulting language—Φ, α, D_{KL}, envelope, separation—allows rigorous cross-domain transfer of insights that have previously been siloed.

Keywords: hierarchical control, autonomy, safe reinforcement learning, constrained MDPs, antifragility, envelope governance, curriculum learning, multi-agent systems

1 Introduction

1.1 The Phenomenon

In any hierarchical system where a senior agent S assumes responsibility for the safety of a subordinate agent J, a temptation arises: S can reduce J's exposure to risk by filtering the environmental states J encounters. This strategy is locally rational. S has superior environmental knowledge and a legitimate mandate to prevent catastrophic outcomes. Within S's optimisation horizon, intervention succeeds: immediate risk decreases.

But across a longer horizon, a pathology emerges. The subordinate agent, deprived of the distributional richness it needs to build an adequate environmental model, becomes fragile: competent within the filtered environment, incompetent outside it. This is not a failure of S's judgment or competence. S is optimising correctly under its own objective function. The problem is structural: S and J are optimising over different time horizons, and the resulting tension cannot be resolved by better calibration of either agent's preferences alone. It requires a change in the *control architecture*.

1.2 Prior Work Across Domains

This phenomenon has been described independently in at least six literatures, each with its own vocabulary:

Table 1: The autonomy-suppression phenomenon across domains.

Domain	Key Concepts	Core Insight	Representative Work
Developmental psychology	Scaffolding, ZPD, attachment	Autonomy requires graded withdrawal of support	Vygotsky [1978], Bowlby [1969]
Organisational theory	Double-loop learning, delegation	Micromanagement suppresses initiative	Argyris [1977], Senge [1990]
Antifragility	Volatility deprivation, iatrogenics	Suppressing stressors increases fragility	Taleb [2012]
Safe RL / constrained MDPs	Shielding, safety sets, CMDPs	Safety constraints reduce exploration	García & Fernández [2015], Chow et al. [2018], Dalal et al. [2018]
Autonomous driving	Safe envelopes, shared control	Boundary control > trajectory control	Erlien et al. [2016], Gerdes & Thornton [2001]
Curriculum learning	Difficulty scheduling, teacher–student	Progressive difficulty improves learning	Bengio et al. [2009], Graves et al. [2017]

Each of these literatures has produced deep domain-specific results. Our aim is not to supersede them but to show that they are studying the same underlying structure. The common thread is a parameterisable intervention that filters the subordinate's experience distribution, and a threshold beyond which filtering degrades long-run performance. Once this structure is recognised, insights transfer: results from safe RL inform parenting theory; results from developmental psychology inform AI alignment; results from autonomous driving inform organisational design.

1.3 Contribution

Our contribution is a *unifying formal vocabulary* consisting of five elements:

(i) the intervention operator Φ parameterised by strength α;
(ii) the KL divergence between the true and filtered distributions as a measure of informational distortion;
(iii) the critical threshold α^* beyond which protection becomes counterproductive;
(iv) the envelope governance architecture as an alternative to trajectory control;
(v) the separation principle as a necessary condition for autonomy development.

We do not claim that any individual result is novel within its home domain. We claim that the *unification* is novel, and that the formal language enables cross-domain transfer that was previously unavailable.

2 System Model

2.1 Environment and Agents

Let the environment be a finite MDP $(\mathcal{X}, \mathcal{A}, T, R, \gamma)$ with state space $\mathcal{X}$ ($|\mathcal{X}| < \infty$), action space $\mathcal{A}$ ($|\mathcal{A}| < \infty$), transition kernel $T : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$, reward function $R : \mathcal{X} \times \mathcal{A} \to [0, R_{\max}]$, and discount factor $\gamma \in [0, 1)$. We assume T induces an ergodic Markov chain under any policy with full support.

Definition 2.1 (Senior Agent)**.** $S = (\pi^S, U^S, \gamma^S)$, where π^S is S's intervention policy mapping states to intervention decisions, U^S is S's utility function (encoding risk-aversion), and $\gamma^S \in [0, 1)$ is S's temporal discount factor.

Definition 2.2 (Junior Agent)**.** $J = (\pi^J, U^J, \gamma^J, \hat{M}, \mathcal{L})$, where $\hat{M}$ is J's learned environment model and $\mathcal{L}$ is a fixed learning algorithm (e.g., Q-learning) satisfying standard convergence conditions (Robbins–Monro step sizes, infinite visitation of all state–action pairs in the accessible set). The assumption of a fixed $\mathcal{L}$ is a limitation: adaptive learners (meta-learning, curiosity-driven exploration) may partially compensate for distributional restriction. We discuss this in Section 12.

2.2 The Intervention Operator

Previous versions of this manuscript modelled Φ as an arbitrary measurable map $\Delta(\mathcal{X}) \to \Delta(\mathcal{X})$, which yielded excessive generality. Following the shared-control literature [Erlien et al., 2016], we restrict attention to a specific parametric family.

Definition 2.3 (Truncation Intervention)**.** The intervention operator Φ_α with strength $\alpha \in [0, 1]$ acts by removing the fraction α of states with highest risk $R_{\text{risk}}(x)$ and renormalising:

$$P^\Phi(x) = P(x \mid x \in \mathcal{X}_\alpha), \quad \text{where} \quad \mathcal{X}_\alpha = \{x \in \mathcal{X} : R_{\text{risk}}(x) \leq r_\alpha\} \tag{1}$$

with r_α chosen so that $|\mathcal{X}_\alpha| = \lfloor (1-\alpha)|\mathcal{X}| \rfloor$. At $\alpha = 0$, $\mathcal{X}_0 = \mathcal{X}$ (no intervention). As $\alpha \to 1$, $\mathcal{X}_\alpha$ shrinks to the single safest state.

Remark 2.4 (Rationality of S). This truncation model captures the empirically dominant pattern: supervisors remove the most dangerous states first. S is not making an error. Under S's own discount-weighted risk objective (Equation (2) below), higher α is strictly preferred. The conflict arises not from irrationality but from the structural fact that S and J optimise over different horizons. This observation—that the pathology is architectural rather than attributable to poor judgment—is the philosophical core of the framework.

Remark 2.5 (Scope of the model). Truncation is the simplest intervention class that captures support reduction. Real-world interventions may also include soft shielding (penalty-based; Dalal et al., 2018), action masking [García & Fernández, 2015], or Lyapunov-based constraint enforcement [Chow et al., 2018]. We treat extensions in Section 12.

2.3 The Horizon Conflict

S minimises short-horizon cumulative risk:

$$V^S(\alpha) = -\sum_{t=0}^{h} (\gamma^S)^t \cdot \mathbb{E}\big[R_{\text{risk}}(x_t) \mid \Phi_\alpha\big] \tag{2}$$

J requires long-horizon adaptive competence measured under the *true* distribution:

$$V^J(\pi^J) = \sum_{t=0}^{H} (\gamma^J)^t \cdot \mathbb{E}_P\big[R(x_t, a_t) \mid \pi^J\big] \tag{3}$$

where $H \gg h$ and typically $\gamma^J > \gamma^S$. The key observation is that $\partial V^S / \partial \alpha > 0$ (more intervention reduces short-run risk) while, beyond a threshold, $\partial V^J / \partial \alpha < 0$ (more intervention reduces long-run competence). This tension is not resolvable by preference adjustment; it is a structural property of the architecture. In economic terms, it is a multi-principal problem with misaligned time preferences.

3 Informational Degradation Under Filtering

3.1 Support Coverage and Reward-Structure Preservation

The quality of J's learned model $\hat{M}$ depends on two separable properties of the filtered distribution.

Definition 3.1 (Support Coverage)**.** The coverage ratio is $c(\alpha) = |\mathcal{X}_\alpha|/|\mathcal{X}| = 1 - \alpha$. Full coverage ($c = 1$) means no states are filtered.

Definition 3.2 (Reward-Structure Preservation)**.** The filtered distribution P^Φ preserves the reward structure of the full MDP if for all policy pairs π, π' with support in $\mathcal{X}_\alpha$:

$$\mathbb{E}_{P^\Phi}[V^\pi] \geq \mathbb{E}_{P^\Phi}[V^{\pi'}] \iff \mathbb{E}_P[V^\pi] \geq \mathbb{E}_P[V^{\pi'}].$$

That is, the relative ordering of policy values is preserved under filtering.

Remark 3.3. These are independent conditions. An intervention can have high coverage but distort value ordering (e.g., if it removes states that make a suboptimal policy look deceptively good). Conversely, an intervention with low coverage can preserve reward structure if the removed states contribute only noise.

3.2 Regret Under Distributional Shift

The following result connects intervention strength to policy quality. It is a direct application of the simulation lemma [Kearns & Singh, 2002] and Pinsker's inequality, and is stated as a proposition to reflect its derivative character.

Proposition 3.4 (Regret Upper Bound Under Filtering)**.** *Let π^* be the optimal policy under P and $\hat{\pi}$ the policy learned under P^Φ. Assume the value function $V^\pi(x)$ is L-Lipschitz in the state distribution under total variation distance. Then cumulative regret over horizon H satisfies:*

$$\text{Regret}_H(\hat{\pi}) \leq H \cdot L \cdot d_{\text{TV}}(P, P^\Phi) \leq H \cdot L \cdot \sqrt{\tfrac{1}{2} D_{\text{KL}}(P \,\|\, P^\Phi)} \tag{4}$$

Proof. The first inequality is the simulation lemma: the value difference under two distributions is bounded by the Lipschitz constant of the value function times the total variation distance between the distributions. The second inequality is Pinsker's: $d_{\text{TV}}(P, Q) \leq \sqrt{\frac{1}{2} D_{\text{KL}}(P \| Q)}$. Note the direction: Pinsker bounds TV *from above* by KL, yielding an *upper* bound on regret. □

Remark 3.5. The bound is loose but directionally informative: it confirms that increasing α increases the regret ceiling. For tighter bounds under specific value-function classes, see Csiszár & Shields [2004] on f-divergence families, or Munos [2003] on Bellman-residual-based bounds.

Corollary 3.6. *As $\alpha \to 1$, $D_{\text{KL}}(P \| P^\Phi) \to \infty$ (since P^Φ assigns zero mass to states in $\mathcal{X} \setminus \mathcal{X}_\alpha$), and the regret bound diverges. In practical terms, the learned policy becomes arbitrarily unreliable under deployment in the full environment.*

4 Fragility and the Critical Threshold α^*

4.1 Fragility as Transfer Loss

Following Taleb [2012] and the distributional robustness literature [Ben-Tal et al., 2009], we define fragility as the gap between potential and realised performance under the true distribution.

Definition 4.1 (Fragility)**.**

$$F(\alpha) = \mathbb{E}_P[V^{\pi^*}] - \mathbb{E}_P[V^{\hat{\pi}_\alpha}]$$

where π^* is the optimal policy under P and $\hat{\pi}_\alpha$ is the policy learned under Φ_α. Note $F(\alpha) \geq 0$ with equality iff filtering does not affect optimal policy selection.

4.2 Formal Characterisation of α^*

A central concept in the framework is the threshold below which intervention is benign.

Definition 4.2 (Critical Threshold)**.** $\alpha^* = \sup\{\alpha \in [0,1] : F(\alpha) = 0\}$. That is, α^* is the maximum intervention strength that does not degrade the optimal policy.

Proposition 4.3 (Existence and Characterisation of α^*)**.** *In a finite MDP with a unique optimal policy π^* and risk-ordered state removal (Definition 2.3):*

(i) $\alpha^* > 0$ *whenever there exist states with $R_{\text{risk}}(x)$ above the removal threshold that have zero visitation probability under π^*;*

(ii) $\alpha^* < 1$ *whenever π^* visits at least two distinct states (i.e., the environment is non-trivial);*

(iii) α^* *is determined by the structure of the MDP: specifically, $\alpha^* = |\{x \in \mathcal{X} : x \notin \text{supp}(d^{\pi^*})\}|/|\mathcal{X}|$, where d^{π^*} is the stationary state distribution under π^*.*

Proof. (i) Removing states with zero visitation under π^* does not alter the Bellman equation at any visited state, so the optimal policy on $\mathcal{X}_\alpha$ agrees with π^* on $\mathcal{X}_\alpha$, giving $F(\alpha) = 0$. (ii) If π^* visits at least two states, the set of zero-visitation states is a strict subset of $\mathcal{X}$, so eventually removal reaches visited states. (iii) Follows from the construction: removal is ordered by risk, and the first policy-disrupting removal occurs when a visited state is removed. □

Remark 4.4. The value of α^* is generally not identifiable from data without knowledge of the full MDP structure, because J never observes the states that S has removed. In practice, α^* must be estimated conservatively—e.g., by monitoring J's transfer performance as α is gradually reduced. This connects to the problem of off-policy evaluation in offline RL [Levine et al., 2020].

4.3 Monotonicity of Fragility Beyond α^*

Proposition 4.5 (Fragility Monotonicity)**.** *Under the conditions of Proposition 4.3, for $\alpha > \alpha^*$, $F(\alpha)$ is non-decreasing in α. Strict monotonicity holds whenever each increment in α removes at least one state on an optimal trajectory.*

Proof sketch. For $\alpha > \alpha^*$, any further removal hits states visited by π^*. At each such removal, the Bellman equation at predecessor states must be resolved with a restricted action/transition set, which (by the optimality of π^*) cannot improve the value. The fragility increment is non-negative; it is strictly positive whenever the removed state was the unique successor under an optimal action. □

Remark 4.6. An important qualification: in environments where multiple states on optimal trajectories are fungible (identical reward and transition structure), removing one may not increase fragility. The strict monotonicity condition requires that each removed state is non-redundant. This nuance is invisible in the informal claim "more protection = more fragility," which is why formal conditions matter.

4.4 Dynamical Systems Analogy (Informal)

The formal results above have an evocative dynamical-systems analog. Consider J's competence z evolving as $z_{t+1} = f(z_t, x_t) + \varepsilon_t$. When perturbations ε are suppressed, the system may converge to a locally stable but globally suboptimal attractor—a metastable equilibrium from which J cannot escape without perturbation. This analogy is suggestive rather than formal, and we do not derive results from it. We include it because it connects to the antifragility intuition [Taleb, 2012] and to the concept of local optima in non-convex optimisation.

5 Exploration Suppression Under Intervention

The exploration–exploitation trade-off is classical in RL [Sutton & Barto, 2018, Auer et al., 2002]. We now show how α distorts this trade-off in the hierarchical setting.

Definition 5.1 (Effective Exploration Rate)**.** Given agent J with intrinsic exploration rate η_0 (e.g., the ε in ε-greedy) and intervention Φ_α, the effective exploration rate is:

$$\eta_{\text{eff}}(\alpha) = \eta_0 \cdot c(\alpha) = \eta_0 \cdot (1 - \alpha) \tag{5}$$

reflecting the fact that exploration can only discover states within $\mathcal{X}_\alpha$.

Remark 5.2. This is a coarse proxy. In practice, effective exploration depends not just on the *size* of the accessible state space but on its *structure*—whether decision-relevant states (near value-function discontinuities, reward boundaries, bottleneck transitions) are retained. The metric captures the aggregate effect but not the geometry of the restriction. For structure-aware exploration bounds, see Jaksch et al. [2010] on UCRL.

Proposition 5.3. *For $\alpha > 0$, $\eta_{\text{eff}} < \eta_0$. As $\alpha \to 1$, $\eta_{\text{eff}} \to 0$. In the limit, J's policy converges to a greedy policy over the residual state space, with no capacity to discover alternative strategies.*

This is formally equivalent to the problem of insufficient coverage in off-policy RL [Levine et al., 2020, Kumar et al., 2020] and connects to the well-known result that constrained MDPs can produce arbitrarily suboptimal policies when safety constraints are overly restrictive [García & Fernández, 2015].

6 Objective Displacement

Under persistent high-α intervention, a secondary pathology emerges: J begins optimising for S's approval signal rather than for environmental competence. This is well-documented as reward hacking [Amodei et al., 2016], specification gaming [Krakovna et al., 2020], Goodhart's law [Manheim & Garrabrant, 2019], teaching to the test [Koretz, 2002], and goal displacement [Merton, 1957]. Our contribution here is not the observation but the claim that it is *endogenous to high-α architectures* rather than requiring agent malice or design error.

Definition 6.1 (Objective Displacement)**.** Let $r_{\text{env}}(x, a)$ be the environmental reward and $r_S(x, a)$ the approval signal from S. Objective displacement occurs when J's effective reward function shifts from r_{env} toward r_S:

$$\hat{r}(x, a) = (1 - \lambda) \cdot r_{\text{env}}(x, a) + \lambda \cdot r_S(x, a) \tag{6}$$

for $\lambda \in [0, 1]$. When $\lambda \approx 1$, the agent is reward-hacking.

Conjecture 6.2 (Endogenous Displacement)**.** *Under persistent high-α intervention, λ increases over time. Mechanistically: when J's accessible state space is dominated by S's intervention boundary, the most predictive signal for J's future states becomes S's approval, not the environmental reward. J therefore learns to attend to r_S because it is the most informationally efficient signal in the restricted environment.*

Remark 6.3. We state this as a conjecture rather than a proposition because a formal proof requires specifying J's internal credit-assignment mechanism. The claim is supported by the reward-shaping literature [Ng et al., 1999]: auxiliary reward signals that dominate the primary signal in frequency and predictive value will dominate the learned policy. A connection to causal inference is also possible: Φ performs a causal intervention on the state distribution [Pearl, 2009], and J's learning algorithm may confound S's approval with genuine environmental structure. We leave formalisation to future work.

7 Distributed Supervision

We consider the effect of distributing the supervisory function across multiple agents $\{S_1, \ldots, S_k\}$ with intervention operators $\Phi_1, \ldots, \Phi_k$.

Definition 7.1 (Convex-Combination Aggregation)**.** The aggregate filtered distribution under convex combination is $P^{\Phi_{\text{agg}}} = \sum_i w_i \cdot P^{\Phi_i}$, with weights $w_i > 0$ and $\sum w_i = 1$.

Proposition 7.2 (Divergence Reduction Under Convex Aggregation)**.** *Under convex-combination aggregation, if the supervisors have non-identical safe sets ($\mathcal{X}_{\alpha_i} \neq \mathcal{X}_{\alpha_j}$ for some i, j), then:*

$$D_{\text{KL}}(P \| P^{\Phi_{\text{agg}}}) \leq \sum_i w_i \cdot D_{\text{KL}}(P \| P^{\Phi_i}) \tag{7}$$

In particular, $D_{\text{KL}}(P \| P^{\Phi_{\text{agg}}}) \leq \max_i D_{\text{KL}}(P \| P^{\Phi_i})$.

Proof. By the joint convexity of KL divergence in its second argument: $D_{\text{KL}}(P \| \sum_i w_i Q_i) \leq \sum_i w_i D_{\text{KL}}(P \| Q_i)$. □

Remark 7.3. The stronger claim $D_{\text{KL}}(P \| P^{\Phi_{\text{agg}}}) \leq \min_i D_{\text{KL}}(P \| P^{\Phi_i})$ does *not* hold in general under convex combination. It does hold in the special case where aggregation is by union of safe sets ($\mathcal{X}_{\text{agg}} = \bigcup_i \mathcal{X}_{\alpha_i}$), since union strictly increases support. The practical implication is that multiple mentors with diverse safe-set specifications produce a richer training distribution than any single mentor, under any reasonable aggregation scheme—an observation consistent with ensemble theory in ML [Dietterich, 2000] and the wisdom-of-crowds literature [Surowiecki, 2004].

8 Observational vs. Declarative Learning

We briefly formalise the distinction between learning from demonstrations and learning from explicit instructions.

Definition 8.1 (Learning Modalities)**.** Declarative learning provides J with a policy mapping $\pi_{\text{declared}} : \mathcal{X} \to \mathcal{A}$. Observational learning provides J with trajectory data $\tau = \{(x_t, a_t, r_t)\}$ from S's interaction with Ω.

Proposition 8.2 (Informational Advantage of Observation)**.** *The mutual information between the trajectory data and the true environment model M^* is at least as great as the mutual information between the declarative policy and M^*:*

$$I(\tau;\, M^*) \geq I(\pi_{\text{declared}};\, M^*) \tag{8}$$

with strict inequality whenever S's policy is stochastic or the environment is stochastic.

Proof sketch. The trajectory data contains the policy as a marginal (the action sequence) but also contains the environmental transitions and reward realisations. By the data processing inequality, $I(\tau; M^*) \geq I(g(\tau); M^*)$ for any function g, and the declarative policy is extractable from the trajectory as a deterministic function. Strict inequality follows because stochasticity in transitions or policy means the trajectory carries information about M^* that the marginal action sequence does not. □

Remark 8.3. This result is framed in terms of mutual information rather than Shannon entropy, because greater entropy does not imply greater usefulness for model construction—high-entropy data can be noisy and uninformative. Mutual information with respect to M^* quantifies how much trajectory data *reduces uncertainty about the environment* specifically. This connects to the imitation learning literature [Ross et al., 2011, Argall et al., 2009].

9 Envelope Governance

9.1 From Trajectory Control to Boundary Constraint

The central architectural proposal of this framework is a shift from *trajectory control* (constraining *what J does*) to *boundary constraint* (constraining *where J can go*). This distinction is not new: it appears in the safe-envelope literature for autonomous driving [Erlien et al., 2016, Gerdes & Thornton, 2001], in constrained MDPs [Altman, 1999], in Lyapunov-based safe RL [Chow et al., 2018], and in recovery policies [Dalal et al., 2018]. Our contribution is to show that this distinction recurs across all the domains in our framework and to give it a unified treatment.

Definition 9.1 (Environmental Envelope)**.** $E \subseteq \mathcal{X}$ is an environmental envelope if: (i) all states in E are survivable: $\forall x \in E$, $L_{\text{cat}}(x) < L_{\text{crit}}$, where L_{cat} is the catastrophic loss function and L_{crit} is the irreversibility threshold; (ii) E is forward-invariant under at least one recovery policy (there exists π_{safe} such that the system can return to E from any state in E).

Definition 9.2 (Envelope Governance)**.** Under envelope governance, S defines and maintains E but does not constrain J's policy within E. The intervention operator is:

$$\Phi^E(P) = P(\cdot \mid x \in E) \tag{9}$$

That is: P conditioned on the envelope, with no further filtering within E. The key distinction: trajectory control constrains J's actions; envelope governance constrains J's reachable states. Within E, J is free to fail, explore, and construct its own model.

9.2 Optimality Conditions

Proposition 9.3 (Envelope Governance Properties)**.** *Let E be the maximal safe subset of $\mathcal{X}$ (Definition 9.1). Then:*

(i) Safety: *no state in E produces catastrophic loss, by construction;*

(ii) Asymptotic optimality within E: *if J's learning algorithm $\mathcal{L}$ satisfies standard convergence conditions and all state–action pairs in E are visited infinitely often (ergodicity within E), then $\hat{\pi} \to \pi^*_E$ as $t \to \infty$, where π^*_E is the optimal policy for the MDP restricted to E;*

(iii) Information maximisation: *since E is maximal, the coverage ratio c is maximised among all safe subsets, which by Proposition 3.4 minimises the regret bound among all safe interventions.*

Remark 9.4. The result is more precisely described as: envelope governance with the maximal safe set achieves the best available trade-off among truncation-class interventions, but it does not achieve global optimality if $E \subsetneq \mathcal{X}$. The agent learns π^*_E, not π^*. Whether π^*_E is a good approximation to π^* depends on how much of the value-relevant state space lies inside E. Convergence *rates* (not just asymptotic convergence) are critical for practical relevance—see Jaksch et al. [2010] for finite-time bounds in ergodic MDPs.

9.3 Progressive Envelope Expansion

In practice, the envelope should expand as J's competence grows. This idea unifies several independently developed concepts:

Definition 9.5 (Envelope Schedule)**.** An envelope schedule is a nested sequence $\{E_0 \subseteq E_1 \subseteq \cdots \subseteq \mathcal{X}\}$ where expansion from E_t to E_{t+1} is triggered when J achieves competence threshold $C_{\min}$ on E_t (measured as, e.g., average return within δ of $\pi^*_{E_t}$).

The schedule converges to $\mathcal{X}$ (full autonomy) if J demonstrates competence at each stage. If J fails to reach $C_{\min}$, the envelope holds or contracts. This is precisely the zone of proximal

Table 2: Progressive expansion across domains.

Domain	Concept	Formal Equivalent
Developmental psychology	Zone of proximal development	$E_{t+1} = E_t \cup \{\text{states within } J\text{'s competence } +\delta\}$
Education	Scaffolding	Temporary supports removed as competence demonstrated
Curriculum learning (ML)	Difficulty scheduling	Training on $E_0 \subset E_1 \subset \cdots \subset \mathcal{X}$
Autonomous driving	ODD expansion	Validated safe envelope grows with demonstrated capability
Organisational theory	Progressive delegation	Authority scope expands with track record

development [Vygotsky, 1978] expressed in control-theoretic terms: the envelope at time t represents the set of challenges that are within J's current capacity plus a stretch zone. The fact that developmental psychology, curriculum learning, autonomous driving, and organisational delegation all converge on the same scheme—despite being developed independently—is evidence that the abstract structure captures something real.

9.4 Illustrative Example: Cliff Gridworld

We illustrate the framework with a standard Cliff Walking environment [Sutton & Barto, 2018, Example 6.6], modified to show the effect of varying α.

Setup. A 4×12 grid. Start state: bottom-left. Goal: bottom-right. The bottom row (excluding start and goal) consists of cliff states with reward -100 and episode termination. All other transitions yield reward -1. The optimal policy follows the bottom row adjacent to the cliff (shortest path, highest risk).

Intervention regimes. We compare three levels of α:

Table 3: Cliff Gridworld under different intervention regimes.

Regime	Accessible States	Outcome
$\alpha = 0$ (none)	All 48 states	Agent discovers optimal cliff-adjacent path after some falls. High training cost, optimal deployment.
$\alpha \approx 0.25$ (envelope)	All except cliff cells (36 states)	Agent cannot fall off cliff. Learns near-optimal safe path (one row above cliff). Zero catastrophic failures, good deployment.
$\alpha \approx 0.75$ (heavy)	Only top 3 rows (12 states)	Agent learns to traverse top rows. Very long path. No model of lower grid. Catastrophic if deployed near cliff.

In this example, $\alpha^* \approx 0.25$: removing the cliff states does not alter the safe-optimal policy. Beyond α^*, each additional row removed degrades J's model of the lower grid and forces increasingly suboptimal paths.

Envelope governance corresponds to $\alpha \approx 0.25$: the cliff cells are removed (safety), but the rest of the grid is accessible (learning). **Progressive expansion** would first restrict to the top half, then expand downward as J demonstrates competence, eventually opening the cliff-adjacent row once J has learned reliable navigation.

10 The Separation Principle

The results above imply a necessary condition for the development of autonomous agents in hierarchical systems.

Proposition 10.1 (Necessity of Control Withdrawal)**.** *Let J be a junior agent with a convergent learning algorithm $\mathcal{L}$ in a finite ergodic MDP. If the intervention strength satisfies $\alpha(t) > \alpha^*$ for all $t > t_0$ (persistent over-intervention), then $\hat{\pi}$ converges to $\pi^*_{\mathcal{X}_\alpha}$, which is strictly dominated by π^*_E (the optimal policy over the maximal safe envelope $E \supseteq \mathcal{X}_\alpha$).*

Proof. Under persistent $\alpha > \alpha^*$, J's learning converges (by assumption on $\mathcal{L}$) to the optimal policy for the restricted MDP on $\mathcal{X}_\alpha$. Since $\alpha > \alpha^*$, by Definition 4.2, this policy is strictly suboptimal on $\mathcal{X}_\alpha$ evaluated under the true distribution P. The envelope E contains $\mathcal{X}_\alpha$ and additional safe states, so π^*_E has access to a richer transition structure and achieves higher expected return. Therefore $\pi^*_{\mathcal{X}_\alpha}$ is strictly dominated. □

Remark 10.2. An important caveat: in environments where the catastrophic states are dense and the safe set is small, persistent intervention may be *necessary*—the cliff-without-a-path scenario. The separation principle applies only when the maximal safe envelope is strictly larger than the intervened set, i.e., when S is being more restrictive than safety requires.

11 Cross-Domain Instantiation

The formal vocabulary maps onto concrete domains as follows:

Table 4: Cross-domain instantiation of the framework.

Domain	**S**	**J**	**Envelope E**	**$\alpha > \alpha^*$ Manifests As**
Parenting	Parent	Child	Safe physical / social environment	Helplessness, anxiety, fragility
Education	Curriculum	Student	Scaffolded problem sets	Teaching to the test, surface learning
Organisation	Manager	Employee	Delegated authority scope	Learned helplessness, low initiative
AI Training	RLHF / overseer	Agent	Constrained action space	Reward hacking, specification gaming
Driving	Safety system	AV planner	Safe driving envelope	Overly conservative, unable to merge
Governance	Central authority	Local actors	Legal / institutional framework	Regulatory capture, stagnation

The value of the unified framework is that insights transfer. The safe-driving literature's envelope concept applies to parenting. The reward-hacking literature's objective displacement applies to organisational management. The curriculum-learning literature's progressive scheduling applies to governance. These transfers become precise rather than metaphorical because they share the same formal structure: Φ, α, D_{KL}, E, separation.

12 Limitations and Extensions

Fixed learning algorithm. The framework assumes J uses a fixed learning algorithm $\mathcal{L}$. If J employs meta-learning [Finn et al., 2017], curiosity-driven exploration [Pathak et al., 2017], or other adaptive mechanisms, it may partially compensate for distributional restriction by up-weighting novel states. Several of our results (particularly Proposition 5.3 on exploration suppression) may not hold for such agents.

Truncation model. The restriction to truncation interventions (Definition 2.3) excludes soft-shielding and reward-penalty approaches. In soft-shielding [Dalal et al., 2018], dangerous states are penalised but not removed, which preserves support coverage while reducing visitation frequency. The informational consequences are different: the training distribution is *reweighted* rather than *truncated*, and the relevant divergence measure shifts from support-sensitive KL to density-ratio-based measures. Extending the framework to soft interventions is natural but would require replacing Proposition 3.4 with importance-weighted bounds.

Computability of the maximal envelope. Definition 9.1 requires identifying the maximal safe subset of $\mathcal{X}$, which may be NP-hard in large state spaces. In practice, envelopes are often specified by domain experts (as in autonomous driving) or learned via model-based RL with safety constraints [Berkenkamp et al., 2017]. The relationship between envelope accuracy and agent performance is an open question.

Partial observability. The framework assumes full state observability. Under partial observability (POMDPs), J cannot distinguish between states it has never visited and states obscured by noisy observation. This makes the informational-sufficiency analysis considerably more complex and is deferred to future work.

Multi-supervisor dynamics. Section 7 treats distributed supervision as a static aggregation problem. In practice, supervisors interact and may coordinate or compete, leading to game-theoretic dynamics not captured by convex combination. The credit-assignment problem—which supervisor's safe set is most informative—is also open.

End-to-end learning of the envelope schedule. Definition 9.5 specifies the envelope schedule exogenously. An attractive extension is joint learning of the schedule and the agent's policy, connecting to the automatic curriculum generation literature [Graves et al., 2017, Portelas et al., 2020].

13 Conclusion

This paper has developed a formal vocabulary for analysing autonomy suppression in hierarchical multi-agent systems. The vocabulary consists of five elements: the parameterised intervention operator Φ_α; the KL divergence between true and filtered distributions as a measure of informational distortion; the critical threshold α^* beyond which intervention degrades agent quality; the envelope governance architecture as a principled alternative to trajectory control; and the separation principle as a necessary condition for autonomy.

The framework's primary contribution is *unification*. The phenomena of distributional shift, exploration suppression, reward hacking, and volatility-induced fragility have been studied in depth within their respective domains. What was missing was a shared formal language showing that these are manifestations of a single structural parameter—α—operating in a single architectural pattern—the hierarchy with horizon conflict. Once this structure is recognised, results and interventions transfer across domains.

One philosophical point deserves emphasis. The senior agent S is not the villain of this framework. S is optimising rationally under its own objective function. The pathology is structural, not moral: it emerges from the architecture of the hierarchy, specifically from the misalignment of optimisation horizons. The solution is equally structural: replace trajectory control with envelope governance, expand the envelope progressively, and recognise that the discomfort of observed failure within the safe boundary is the necessary informational cost of producing an agent capable of autonomous function beyond it.

References

Altman, E. (1999). *Constrained Markov Decision Processes*. Chapman & Hall/CRC.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv:1606.06565*.

Anderson, S. J., Peters, S. C., Pilutti, T. E., & Iagnemma, K. (2012). An optimal-control-based framework for trajectory planning, threat assessment, and semi-autonomous control of passenger vehicles in hazard avoidance scenarios. *International Journal of Vehicle Autonomous Systems*, 10(1/2), 1–60.

Argall, B. D., Chernova, S., Veloso, M., & Browning, B. (2009). A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5), 469–483.

Argyris, C. (1977). Double loop learning in organizations. *Harvard Business Review*, 55(5), 115–125.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3), 235–256.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, 79(1–2), 151–175.

Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009). *Robust Optimization*. Princeton University Press.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of ICML*.

Berkenkamp, F., Turchetta, M., Schoellig, A. P., & Krause, A. (2017). Safe model-based reinforcement learning with stability guarantees. In *Advances in NeurIPS*.

Bowlby, J. (1969). *Attachment and Loss, Vol. 1: Attachment*. Basic Books.

Chow, Y., Ghavamzadeh, M., Janson, L., & Pavone, M. (2018). Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(1), 6070–6120.

Csiszár, I., & Shields, P. C. (2004). Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4), 417–528.

Dalal, G., Dvijotham, K., Vecerik, M., Hester, T., Paduraru, C., & Tassa, Y. (2018). Safe exploration in continuous action spaces. *arXiv:1801.08757*.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems*, pp. 1–15. Springer.

Erlien, S. M., Fujita, S., & Gerdes, J. C. (2016). Shared steering control using safe envelopes for obstacle avoidance and vehicle stability. *IEEE Transactions on Intelligent Transportation Systems*, 17(2), 441–451.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of ICML*.

García, J., & Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1), 1437–1480.

Gerdes, J. C., & Thornton, S. M. (2001). Implementable ethics for autonomous vehicles. In *Autonomes Fahren*, Springer.

Graves, A., Bellemare, M. G., Menick, J., Munos, R., & Kavukcuoglu, K. (2017). Automated curriculum learning for neural networks. In *Proceedings of ICML*.

Jaksch, T., Ortner, R., & Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11, 1563–1600.

Kearns, M. J., & Singh, S. P. (2002). Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2–3), 209–232.

Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources*, 37(4), 752–777.

Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., & Legg, S. (2020). Specification gaming: the flip side of AI ingenuity. *DeepMind Blog*.

Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. In *Advances in NeurIPS*.

Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv:2005.01643*.

Manheim, D., & Garrabrant, S. (2019). Categorizing variants of Goodhart's law. *arXiv:1803.04585*.

Merton, R. K. (1957). *Social Theory and Social Structure*. Free Press.

Munos, R. (2003). Error bounds for approximate policy iteration. In *Proceedings of ICML*.

Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of ICML*.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-predictive next feature. In *Proceedings of ICML*.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.

Portelas, R., Colas, C., Weng, L., Hofmann, K., & Oudeyer, P.-Y. (2020). Automatic curriculum learning for deep RL: A short survey. In *Proceedings of IJCAI*.

Ross, S., Gordon, G., & Bagnell, D. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*.

Senge, P. M. (1990). *The Fifth Discipline*. Doubleday.

Surowiecki, J. (2004). *The Wisdom of Crowds*. Doubleday.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press.

Taleb, N. N. (2012). *Antifragile: Things That Gain from Disorder*. Random House.

Vygotsky, L. S. (1978). *Mind in Society*. Harvard University Press.

A Evolutionary Systems as Envelope Governance

A.1 Structural Mapping

The formal framework developed in the main paper characterises a hierarchical architecture in which a senior system defines boundary conditions on a subordinate agent's state space without constraining the agent's trajectory within those boundaries. We now show that biological evolution instantiates this architecture at the population level, with natural selection operating as the intervention operator Φ and the survivability landscape functioning as the environmental envelope E.

Let the state space $\mathcal{X}$ represent the space of possible phenotypes (or, at finer resolution, genotypes). Let the transition kernel T represent the stochastic processes of mutation, recombination, and epigenetic variation that generate novel phenotypic states from existing ones. Let the reward function R represent reproductive fitness.

Natural selection does not specify which phenotypes should arise. It does not guide the trajectory of any individual organism or lineage through phenotype space. It operates exclusively as a filter: phenotypes that fall outside the survivability envelope—those incompatible with thermodynamic maintenance, ecological viability, or reproductive success—are removed from the population over generational time. All phenotypes *within* the envelope persist and propagate with probability proportional to their fitness.

This is precisely the truncation intervention of Definition 2.3:

$$P^{\Phi}(x) = P(x \mid x \in \mathcal{X}_{\alpha})$$

where $\mathcal{X}_{\alpha}$ is the set of viable phenotypes under the current environmental regime. The "intervention strength" α corresponds to the stringency of the selective environment: a harsh environment with narrow viability margins corresponds to high α; a permissive environment with broad viability margins corresponds to low α.

A.2 Absence of Trajectory Control

A defining feature of the evolutionary process is that no mechanism exists for directing the content of genetic variation. Mutations are not goal-directed. Recombination is stochastic. Horizontal gene transfer, transposon activity, and epigenetic modification introduce variation without reference to adaptive utility. The generation of novelty is, in the language of the main paper, *free exploration within the envelope*.

This stands in contrast to hypothetical evolutionary architectures that have been proposed and rejected. *Lamarckian inheritance* posits trajectory control: the organism's acquired adaptations are transmitted to offspring, effectively allowing the environment to guide the direction of variation. This corresponds to low-α intervention with trajectory constraint—precisely the architecture the main paper identifies as suppressive of autonomous adaptation. *Orthogenesis* posits an internal directional force guiding evolution along predetermined lines, corresponding to the senior agent controlling J's policy directly. Both architectures have been empirically falsified. The architecture that persists—undirected variation filtered by selective boundaries—is envelope governance.

A.3 Volatility Deprivation and Evolutionary Fragility

The main paper's Proposition 4.5 (fragility monotonicity) predicts that reducing environmental perturbation below the critical threshold α^* produces agents that are competent within the filtered environment but fragile under deployment in the full state space. Evolutionary biology provides extensive empirical confirmation.

Genetic load under relaxed selection. When selective pressures are artificially relaxed (corresponding to a decrease in the effective α experienced by the population), deleterious mutations accumulate because the boundary that would have filtered them is absent. This is Muller's ratchet in asexual populations [Muller, 1964] and the mutational load problem more generally. The population's average fitness, evaluated against the *full* environmental distribution P rather than the filtered distribution P^{Φ}, declines.

Ecological naïveté. Populations that evolve in predator-free environments (island ecosystems, deep cave systems) rapidly lose antipredator behaviours and morphological defences. When the environmental envelope expands—through introduction of novel predators or connection to mainland ecosystems—these populations experience catastrophic failure. In the language of the framework, their training distribution P^{Φ} lacked support over predator-relevant states, producing a policy $\hat{\pi}$ with zero competence in those regions of the state space.

Monoculture fragility. Agricultural monocultures—populations with artificially reduced genetic diversity, maintained under controlled environmental conditions—are highly productive within the managed envelope but catastrophically vulnerable to perturbations outside it (novel pathogens, climate variation). The Irish Potato Famine (1845–1852) is the canonical example: a population of genetically near-identical *Solanum tuberosum* cultivars, optimised for yield within a stable envelope, collapsed when *Phytophthora infestans* introduced a state outside the training distribution.

A.4 Progressive Envelope Expansion in Evolution

The main paper's envelope schedule (Definition 9.5)—a nested sequence $E_0 \subseteq E_1 \subseteq \cdots \subseteq \mathcal{X}$, expanding as agent competence grows—has a direct evolutionary analog in the pattern of adaptive radiation.

Following mass extinction events, surviving lineages encounter an expanded envelope: ecological niches previously occupied by dominant competitors are vacated. The surviving population, having demonstrated viability in the post-extinction environment (the competence threshold $C_{\min}$), diversifies into the newly accessible state space. The Cambrian explosion, the diversification of mammals following the end-Cretaceous extinction, and the radiation of cichlid fishes in the East African Rift lakes all follow this pattern. In each case, the expansion of the viable state space $E_t \to E_{t+1}$ was not directed by any external agent but occurred through the removal of boundary constraints (dominant competitors, environmental barriers), enabling free exploration in previously inaccessible regions.

A.5 Conclusion

Evolution is a large-scale instance of the envelope governance architecture. Natural selection defines survivability boundaries without controlling trajectories within them. Variation is undirected. Fitness is evaluated under the true distribution. Over-protection (relaxed selection) produces exactly the fragility predicted by the formal framework. The structural equivalence is not metaphorical; it is exact, with the survivability landscape playing the role of the envelope E, selective pressure playing the role of α, and the genotype-phenotype map playing the role of the agent's policy π^J.

B Theological Systems: Non-Intervention and the Problem of Evil

B.1 Preliminary Framing

This appendix treats the theological concept of a divine agent as a problem in hierarchical agency theory. No ontological commitment is required: the analysis is structural. The question is whether the formal properties of hierarchical systems identified in the main paper illuminate a classical problem in philosophical theology—the problem of evil—when the senior agent S is modelled as an omniscient, omnipotent higher-level system and the junior agent J is modelled as a finite agent embedded in a stochastic environment.

The problem of evil, as formalised by Epicurus and elaborated by Hume, Leibniz, Plantinga, and others, asks: if a higher agent possesses both the capacity and the motivation to prevent suffering, why does suffering persist? The standard responses—the free-will defence [Plantinga, 1974], the soul-making theodicy [Hick, 1966], the greater-good argument—are typically framed in moral and metaphysical terms. We offer a systems-theoretic reformulation.

B.2 The Structural Impossibility of Trajectory Control with Autonomy

The main paper's central result is that trajectory control and autonomous adaptation are incompatible objectives in hierarchical systems. Specifically, Proposition 10.1 (Necessity of Control Withdrawal) establishes that persistent intervention above α^* guarantees a strictly dominated agent—one that is less competent than it would have been under envelope governance.

Apply this to the theological case. Model the divine agent as S with:

- Complete environmental knowledge (omniscience): S has access to the full distribution P and the full transition kernel T.
- Unbounded intervention capacity (omnipotence): S can implement any intervention operator Φ, including Φ with $\alpha = 1$ (total trajectory control).
- A utility function U^S that includes the long-horizon autonomy of J as a terminal objective.

If S's objective includes J's autonomous competence—that is, if V^J enters U^S with non-negligible weight—then the framework's results impose a constraint on S's rational strategy:

> S cannot simultaneously maximise J's autonomy and minimise J's exposure to adverse states.

This is not a limitation of S's power. It is a structural property of the optimisation problem itself, analogous to the impossibility of simultaneously minimising variance and bias in statistical estimation, or of simultaneously satisfying all axioms in Arrow's impossibility theorem. The constraint is logical, not physical.

B.3 Laws, Consequences, and Mortality as Envelope

In the theological modelling tradition, the divine agent is characterised as establishing *laws* (physical, moral, consequential) rather than dictating individual actions. This architecture maps directly onto envelope governance (Definition 9.2):

Under this mapping, the divine agent defines the envelope E—the set of states that are survivable (not catastrophically irreversible in an ultimate sense)—and the transition structure T that determines consequences. Within E, agents are free to act, err, suffer consequences, and learn. The divine agent does not intervene to prevent individual failures, because doing so would constitute trajectory control and, by Proposition 10.1, would degrade the agent's autonomous capacity.

Table 5: Mapping theological concepts to the formal framework.

Theological Concept	**Framework Equivalent**
Physical laws	State-space boundaries (Definition 9.1)
Moral law / conscience	Reward signal r_{env}, partially observable
Consequences of action	Transition kernel T
Mortality / finitude	Finite horizon H
Suffering	Negative reward states within the envelope

B.4 Suffering as Informational Cost

The main paper establishes (Proposition 3.4) that the regret bound under filtering is proportional to the distributional shift between the training and true distributions. If the divine agent removed all negative-reward states from J's experience (setting α high enough to exclude suffering), then J's model $\hat{M}$ would lack support over adversity-related regions of the state space. The resulting agent would be: incompetent under deployment in the full environment (Corollary 3.6); subject to objective displacement (Conjecture 6.2), optimising for the approval signal of S rather than for genuine environmental understanding; and fragile (Definition 4.1), with high transfer loss when encountering states outside the filtered distribution.

Suffering, in this framing, is not a design failure but an unavoidable informational cost of preserving the conditions under which autonomous adaptation is possible. The relevant states—loss, pain, failure, mortality—are precisely the high-R_{risk} states that a short-horizon optimiser would remove, but whose removal degrades the long-horizon capacity that the system is designed to produce.

B.5 The Horizon Conflict as Theodicy

The main paper identifies the core tension as a misalignment of optimisation horizons: S optimises over horizon h, J requires competence over horizon $H \gg h$. In the theological case, if the divine agent's horizon is taken as infinite (or at least as encompassing J's full developmental trajectory), then the divine agent is *not* a short-horizon optimiser. The divine agent is, by construction, the agent with the longest possible horizon.

This inverts the standard framing. In the main paper, the pathology arises because S is too short-sighted. In the theological case, the claim is that the divine agent's *long* horizon is precisely what explains non-intervention: a truly long-horizon optimiser recognises that short-term suffering is dominated by long-term autonomous competence, and therefore refrains from the trajectory control that a short-horizon optimiser (a finite observer, a grieving agent, a mortal bystander) would demand.

The problem of evil, under this analysis, is not a problem of divine indifference or incapacity. It is a problem of *horizon mismatch between the observer and the architect.* The finite agent, operating under horizon h, perceives non-intervention as neglect. The infinite-horizon architect, operating under horizon H, recognises non-intervention as the only strategy compatible with the terminal objective.

B.6 Boundaries of the Analysis

This analysis does not constitute a theodicy in the traditional sense. It does not claim to resolve the problem of evil morally or metaphysically. It claims only that the formal structure of the problem—a higher agent refraining from intervention despite possessing the capacity to intervene—is not paradoxical when analysed as a hierarchical control problem with competing time horizons. The same structure appears in every domain treated in the main paper; the theological case is distinguished only by the scope of the agents and the horizon of the optimisation.

The analysis is silent on whether such a higher agent exists. It addresses only the conditional: *if* a hierarchical system with these properties is posited, *then* non-intervention is the structurally predicted strategy, not a contradiction.

C Physical Laws as Envelope Governance

C.1 Laws of Physics as Boundary Constraints

A pervasive feature of fundamental physics is that its laws take the form of *constraints on what cannot occur* rather than *prescriptions of what must occur*. This section demonstrates that this structure is formally equivalent to envelope governance as defined in the main paper.

Consider the distinction between two possible architectures for physical law. Under *trajectory control*, laws specify the trajectory of every particle, field, or system at every moment: given initial conditions, the future is uniquely determined, and the laws actively produce the trajectory. Under *boundary constraint*, laws define the boundaries of the accessible state space (the phase space), the conservation laws that restrict transitions, and the causal structure that limits information propagation; within these boundaries, dynamical freedom exists.

Classical mechanics appears to favour the first architecture (deterministic trajectory from initial conditions via Newton's laws). However, reformulations of mechanics—Lagrangian, Hamiltonian, and especially the principle of least action—reveal that the deeper structure is constraint-based.

C.2 The Principle of Least Action as Envelope

The principle of least action states that the trajectory of a physical system between two states is the one that makes the action functional stationary:

$$\delta S = \delta \int L(q, \dot{q}, t)\, dt = 0$$

This is not a prescription of a specific trajectory. It is a *variational boundary condition*: among all kinematically possible trajectories (the full state space $\mathcal{X}$), only those satisfying the stationarity condition are physically realised. The Euler–Lagrange equations derived from this principle are constraints that define the envelope of dynamically permissible trajectories, not instructions that generate them.

The structural parallel is exact:

Table 6: Mapping physical concepts to the formal framework.

Physical Concept	**Framework Equivalent**
Phase space	Full state space $\mathcal{X}$
Conservation laws	Envelope boundary ∂E
Equations of motion	Constraint on transitions within E
Kinematically possible trajectories	Unconstrained exploration space
Dynamically realised trajectories	Trajectories within the envelope

C.3 Conservation Laws as Envelope Boundaries

Each conservation law in physics—conservation of energy, momentum, angular momentum, charge, baryon number, lepton number—defines a hypersurface in phase space that the system cannot leave. These are not trajectory prescriptions; they are boundary constraints. The system is free to evolve in any manner consistent with these constraints.

Noether's theorem [Noether, 1918] establishes that each continuous symmetry of the Lagrangian corresponds to a conserved quantity. The symmetries of the physical laws define the *shape* of the envelope. The dynamics within the envelope are not further constrained by the symmetry itself—the symmetry constrains only the boundary.

This is precisely the architecture of Definition 9.2: the senior system (physical law) defines the envelope E via conservation constraints, and the junior system (the physical degrees of freedom) evolves freely within E.

C.4 Thermodynamic Constraints and the Arrow of Time

The second law of thermodynamics—entropy of an isolated system does not decrease—is a unidirectional boundary constraint. It does not specify *which* microstates the system occupies; it constrains only the macroscopic direction of evolution. Within the constraint, the system explores its accessible microstate space ergodically (the ergodic hypothesis), visiting all microstates compatible with the macroscopic boundary conditions with equal probability.

This is envelope governance with maximal internal freedom: the boundary (non-decreasing entropy) is defined; the trajectory within the boundary is unconstrained and, in the microcanonical ensemble, uniformly distributed.

C.5 Quantum Mechanics: Maximal Freedom Under Minimal Constraint

Quantum mechanics provides perhaps the strongest instance of the envelope governance architecture in fundamental physics.

The Schrödinger equation governs the evolution of the quantum state, but it does not determine measurement outcomes. The Born rule assigns probabilities to outcomes; it does not select among them. The Heisenberg uncertainty relations define boundaries on simultaneous knowledge of conjugate variables—they constrain the *envelope* of measurable states without prescribing which state is realised.

In the path-integral formulation [Feynman, 1948], a quantum system does not follow a single classical trajectory. It explores *all* kinematically possible paths simultaneously, with each path weighted by $\exp(iS/\hbar)$. The classical trajectory emerges as the stationary-phase approximation—the path that dominates the integral—but the fundamental description is one of maximal exploration within the boundary constraints imposed by the action.

This is the formal analog of envelope governance with $\alpha \to 0$ within the envelope: no internal filtering, maximal exploration, with the boundary defined by the action principle and conservation laws.

C.6 Emergence of Complexity

A consequence of the envelope governance architecture in physics is the emergence of complex structure. Because physical laws constrain boundaries but do not control trajectories, the dynamical freedom within the envelope permits the spontaneous formation of structures that are not prescribed by the laws themselves.

Chemistry is not dictated by quantum electrodynamics; it *emerges* from the degrees of freedom that QED leaves unconstrained within the boundary of electromagnetic interaction. Biology is not dictated by chemistry; it emerges from the combinatorial freedom that chemical bonding rules leave unconstrained within the boundary of thermodynamic viability. Intelligence is not dictated by neurobiology; it emerges from the computational degrees of freedom that neural architecture leaves unconstrained within the boundary of metabolic and physical constraint.

At each level of the hierarchy, the higher-level laws define an envelope, and the lower-level dynamics explore freely within it. Complexity arises precisely because the higher level *does not*

control the trajectory of the lower level. If it did—if physical laws prescribed the exact trajectory of every particle—the system would be deterministic, frozen, and incapable of producing emergent structure. The main paper's central result—that trajectory control suppresses the development of autonomous competence—is thus instantiated at the most fundamental level of physical reality.

C.7 Conclusion

The laws of physics are envelope constraints. They define what cannot occur—violations of conservation laws, superluminal causation, entropy decrease in isolated systems—without prescribing what must occur within those boundaries. The resulting architecture is identical in structure to the envelope governance of Definition 9.2: boundary control without trajectory control. The emergence of chemistry, biology, and intelligence from this architecture is the physical analog of the main paper's central prediction: that autonomous, adaptive agents emerge from hierarchical systems only when the senior level governs the boundary rather than the trajectory.

This observation suggests that envelope governance is not merely a useful design principle for artificial systems or a descriptive model of developmental processes. It is a structural feature of hierarchical organisation at every scale—from the laws of physics, through biological evolution, through the contested problem of divine non-intervention, to the engineering of artificial agents. The formal vocabulary of the main paper—Φ, α, D_{KL}, E, separation—provides a language adequate to all of these instantiations.

References

[A1] Darwin, C. (1859). *On the Origin of Species*. John Murray.

[A2] Feynman, R. P. (1948). Space-time approach to non-relativistic quantum mechanics. *Reviews of Modern Physics*, 20(2), 367–387.

[A3] Gould, S. J. (2002). *The Structure of Evolutionary Theory*. Harvard University Press.

[A4] Hick, J. (1966). *Evil and the God of Love*. Harper & Row.

[A5] Muller, H. J. (1964). The relation of recombination to mutational advance. *Mutation Research*, 1(1), 2–9.

[A6] Noether, E. (1918). Invariante Variationsprobleme. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen*, 235–257.

[A7] Plantinga, A. (1974). *God, Freedom, and Evil*. Eerdmans.