Against Latent Fundamentalism

Admissibility Distortion and the Missing Criterion in Representation Learning for Planning

Flyxion
Independent Researcher
github.com/standardgalactic

Abstract

Representation learning for planning lacks a principled criterion for whether a learned projection preserves the distinctions planning requires. We identify that criterion: *admissibility distortion* $D_A(\varphi)$, which equals zero if and only if the projection is admissibility-preserving. The admissibility equivalence relation $\sim_A$—the coarsest relation compatible with adequate planning—is derived from planning requirements alone, not from reward structure, observation statistics, or architectural choices. The central result, the *Observational–Interventional Separation Theorem*, establishes that $\sim_O \not\Rightarrow \sim_I$ in general: this is a structural mismatch between equivalence-generating operations, not an empirical contingency. Three corollaries follow: predictive projections satisfy $D_A(\varphi_P) \not\equiv 0$ in general; compressive projections satisfy $D_A(\varphi_C) \not\equiv 0$; and by the data-processing inequality, any downstream guardrail classifier is bounded above by $I(A; \varphi(X))$, with a Fano-derived Bayes-risk floor that is independent of classifier complexity. A collapse–separation decomposition ($D_A(\varphi) = P(C_\varphi{=}1)\,\mathbb{E}[\Delta_A \mid C_\varphi{=}1]$) distinguishes broad-compressive from catastrophic-blind-spot failure modes. Trajectory-sampling lower bounds enable witnessed violation detection without full dynamics access: any collapsed pair whose reachability sets disagree under a sampled policy proves $D_A(\varphi) > 0$. Admissibility distortion is a foundational target criterion, not yet an engineering metric; the contribution is naming the missing quantity and deriving its properties.

Keywords: representation learning, state abstraction, reachability, admissibility, planning, safety, world models.

1. Introduction

Planning requires choosing among actions by evaluating their consequences. Any representation intended to support planning must therefore preserve, at minimum, the distinctions that determine which futures are accessible. When a representation fails to preserve such distinctions, planning errors follow that cannot be corrected by improving the optimiser, refining the cost function, or adding safety machinery—the failure is logically prior to all of those operations.

Current representation-learning objectives—predictive accuracy, information maximisation, compression progress, contrastive separation—optimise for properties of the observation distribution. None directly optimises for preservation of reachability structure. The question of whether a given latent space is adequate for planning is therefore not addressed by any current training objective; it is at best hoped that predictive or compressive objectives will be admissible incidentally.

We show that the missing criterion is *admissibility distortion* $D_A(\varphi)$. The argument has a fixed logical architecture: planning induces $\sim_A$; $\sim_A$ induces $D_A(\varphi)$; and existing objectives are evaluated by how well they approximate $\mathcal{X}/\sim_A$. The paper's contributions are (i) the derivation of $\sim_A$ and $D_A(\varphi)$ from planning requirements without reference to reward or observations; (ii) the Observational–Interventional Separation Theorem establishing that the gap is structural, not empirical; (iii) three corollaries for prediction, compression, and safety; (iv) a decomposition theorem and trajectory-sampling lower bound for diagnostic use; and (v) a formal asymmetry result showing that positive lower-bound estimates witness genuine violations while zero estimates provide no certificate.

Relation to prior work. The earliest precise statement of the world-model concept appears in Craik (1943), who argued that an organism carrying "a small-scale model of external reality and its own possible actions within its head" can "try out various alternatives" and "react to future situations before they arise." Craik's formulation already identifies what we formalise as reachability structure: the model must represent the agent's *possible actions*, not merely passive observations of the world. The state-abstraction programme (Li et al., 2006; Ravindran & Barto, 2004) asks which collapses are safe for policy quality; admissibility extends this beyond value-function equivalence to reachability-set equivalence. Bisimulation metrics (Ferns et al., 2004; Castro & Precup, 2010; Kemertas & Aumentado-Armstrong, 2021; Zhang et al., 2021) provide continuous relaxations; Proposition 6 shows these are lower bounds on $D_A(\varphi)$. Predictive state representations (Littman et al., 2002) and general value functions (Sutton et al., 2011) encode prediction structure; the separation theorem shows this is insufficient. Pearl's do-calculus (Pearl, 2009) formalises observation vs. intervention; our

theorem is the planning-theoretic analogue. The information bottleneck (Tishby et al., 1999) minimises $I(X;Z)$ subject to $I(Z;Y)$; replacing Y with the admissibility signal yields an admissibility-constrained bottleneck. Gibson's affordance theory (Gibson, 1979) shifts representation from object identity to action possibility, as does $\sim_A$. Model predictive control (Mayne et al., 2000; Garcia et al., 1989) and reachability analysis (Tomlin et al., 2003) assume state representations are already adequate; $D_A(\varphi)$ formalises when that assumption fails. Controllability Gramians and reachability metrics in linear control theory (Kalman, 1960) provide task-specific admissibility criteria under Gaussian dynamics; $\sim_A$ generalises these to arbitrary policy classes without linearity assumptions. Information-state representations in POMDPs (Kaelbling et al., 1998) preserve sufficient statistics for optimal planning under partial observability; extending $\sim_A$ to belief-space reachability connects these frameworks and is a natural direction for future work.

2. Planning Induces Admissibility

2.1. Setup and Reachability

Let $\mathcal{X}$ be a state space, $H \in \mathbb{N}$ a planning horizon, and Π a policy class. Write $x \overset{\pi,h}{\rightsquigarrow} y$ when state y is reachable from x in h steps under policy π. This formalises what Craik (1943) called the capacity to "try out various alternatives" before acting: the reachability set is exactly the set of alternatives available to the agent.

Definition 1 (Reachability Set)**.** $\mathcal{R}_H^\Pi(x) = \{y \in \mathcal{X} : \exists\, \pi \in \Pi,\ x \overset{\pi,H}{\rightsquigarrow} y\}$.

Proposition 1 (Planning Sufficiency Requirement)**.** *If $\mathcal{R}_H^\Pi(x_1) \neq \mathcal{R}_H^\Pi(x_2)$, there exists a planning problem for which the optimal policy from x_1 differs from the optimal policy from x_2. Hence any representation adequate for planning must distinguish states with distinct reachability sets.*

(Proof in Appendix A.)

2.2. Admissibility Equivalence and Distortion

Definition 2 (Admissibility Equivalence)**.** $x_1 \sim_A x_2 \iff \mathcal{R}_H^\Pi(x_1) = \mathcal{R}_H^\Pi(x_2)$. A projection $\varphi : \mathcal{X} \to \mathcal{M}$ is *admissible* if $\varphi(x_1) = \varphi(x_2) \Rightarrow x_1 \sim_A x_2$.

The relation $\sim_A$ is derived, not assumed: it is the coarsest equivalence relation on $\mathcal{X}$ consistent with Proposition 1. This gives it the character of a sufficient statistic (Fisher, 1922): it is the minimally discriminative representation safe for all accessibility-structured planning tasks.

Remark 1 (Capability-relativity). $\sim_A$ is properly written $\sim_A^{(\Pi,H)}$: it depends on the agent's capability class Π and planning horizon H. This is not a deficiency. Every useful equivalence relation in planning is similarly indexed: bisimulation depends on transition structure, observability depends on sensors, controllability depends on actuators. If an agent acquires new capabilities, Π expands and $\sim_A$ may split previously equivalent pairs, requiring finer representation. Admissibility is intentionally relative to a capability class: a representation adequate for one class need not remain adequate for a strictly larger one.

Definition 3 (Admissibility Distortion)**.**

$$D_A(\varphi) \;=\; \mathbb{E}_{x_1,x_2}\big[\mathbf{1}_{\varphi(x_1)=\varphi(x_2)} \cdot d\big(\mathcal{R}_H^\Pi(x_1), \mathcal{R}_H^\Pi(x_2)\big)\big]\,, \tag{1}$$

where $d(\cdot,\cdot)$ is any metric on subsets of $\mathcal{X}$ (e.g. Hausdorff distance) and the expectation is over pairs drawn from the agent's state distribution. $D_A(\varphi) = 0$ iff φ is admissible.

Remark 2 (Metric invariance of the binary question). The question $D_A(\varphi) = 0$ versus $D_A(\varphi) > 0$ is invariant under any choice of d satisfying $d(S,T) = 0 \iff S = T$. The metric affects only the severity weighting assigned to violations, not whether they exist. All theorem-level results (separation, monotonicity, irrecoverable collapse) concern the binary question and are therefore metric-independent.

Theorem 1 (Zero-Distortion Characterisation)**.** $D_A(\varphi) = 0 \iff P(C_\varphi(x_1,x_2) = 1,\ x_1 \not\sim_A x_2) = 0$. *That is, admissibility distortion vanishes iff collapsed pairs are admissibility-equivalent almost surely.*

(Proof in Appendix A.)

Remark 3 (Epistemic status). $D_A(\varphi)$ is mathematically well-defined but not computable in full generality without access to the dynamics in the form needed to enumerate $\mathcal{R}_H^\Pi(x)$. The paper treats $D_A(\varphi)$ as a *target criterion*—the quantity representation learning should minimise—rather than a deployed metric. It plays the role that Kolmogorov complexity plays in compression theory: foundationally clarifying and practically directive without being directly computable in general.

2.3. Decomposition

Define the collapse indicator $C_\varphi(x_1,x_2) = \mathbf{1}_{\varphi(x_1)=\varphi(x_2)}$ and reachability separation $\Delta_A(x_1,x_2) = d(\mathcal{R}_H^\Pi(x_1), \mathcal{R}_H^\Pi(x_2))$.

Proposition 2 (Collapse–Separation Decomposition)**.** $D_A(\varphi) = P(C_\varphi = 1) \cdot \mathbb{E}[\Delta_A \mid C_\varphi = 1]$.

The decomposition distinguishes two failure modes: *broad-compressive* failure (high $P(C_\varphi = 1)$, moderate $\mathbb{E}[\Delta_A \mid C_\varphi = 1]$) and *catastrophic blind-spot* failure (low $P(C_\varphi = 1)$, but the few merged pairs have $\Delta_A \gg 0$). Two representations with equal $D_A(\varphi)$ may call for entirely different interventions.

2.4. Universality and Monotonicity

Theorem 2 (Universality of $\sim_A$)**.** *Let $\mathfrak{T}_A$ be the class of* accessibility tasks*: cost functions depending only on which futures are reachable, not on transition probabilities or reward magnitudes along trajectories. Then $\sim_A = \bigcap_{\mathcal{T}\in\mathfrak{T}_A} \sim_{\mathcal{T}}$,*

where $x_1 \sim_{\mathcal{T}} x_2$ iff the optimal policies from x_1 and x_2 coincide under $\mathcal{T}$. Consequently $\sim_A$ is the largest equivalence relation safe for all tasks in $\mathfrak{T}_A$.

Remark 4 (Scope of universality). Tasks whose costs depend on transition *probabilities* or reward magnitudes (rather than mere reachability) may distinguish states that $\sim_A$ identifies. For such tasks, admissibility must be extended to an *admissibility profile* recording reachable states together with their access-cost distributions; this extension is deferred to future work. The present results establish $\sim_A$ as a necessary condition for all tasks in $\mathfrak{T}_A$, which includes all goal-conditioned and viability objectives.

Theorem 3 (Monotonicity). *If $\sim_{\varphi_1} \subseteq \sim_{\varphi_2}$ (so φ_2 collapses at least every pair φ_1 collapses), then $D_A(\varphi_1) \leq D_A(\varphi_2)$.*

Let $\mathcal{N} = \sim_{\varphi_2} \setminus \sim_{\varphi_1}$ denote newly collapsed pairs.

Corollary 1 (Accounting Identity). $D_A(\varphi_2) - D_A(\varphi_1) = \mathbb{E}[\mathbf{1}_{\mathcal{N}}(x_1, x_2) \cdot \Delta_A(x_1, x_2)]$.

The increase in distortion from any compression step is exactly the expected reachability separation of the newly merged pairs—not a bound, but an identity. This makes the admissibility cost of each compression decision explicit and in principle auditable.

Corollary 2 (Strict Monotonicity). *If there exists a pair $(x_1, x_2) \in \mathcal{N}$ with $\Delta_A(x_1, x_2) > 0$ on a set of positive measure, then $D_A(\varphi_2) > D_A(\varphi_1)$.*

(Immediate from Corollary 1: the expectation is strictly positive when a positive-measure set of newly merged pairs have nonzero separation.) This identifies exactly when compression increases distortion: it does so whenever any newly merged pair consists of states with genuinely different reachable futures.

Proposition 3 (Lipschitz Stability). *Suppose reachability sets are Lipschitz in the state metric: $d_H(\mathcal{R}_H^{\Pi}(x_1), \mathcal{R}_H^{\Pi}(x_2)) \leq L\, d_{\mathcal{X}}(x_1, x_2)$ for some $L > 0$ and state metric $d_{\mathcal{X}}$. Then*

$$D_A(\varphi) \;\leq\; L\, \mathbb{E}\big[\mathbf{1}_{C_\varphi = 1} \cdot d_{\mathcal{X}}(x_1, x_2)\big]\,.$$

(Proof: substitute the Lipschitz bound into Definition 3.) This connects admissibility distortion to geometric clustering: any encoder that collapses only nearby states (small $d_{\mathcal{X}}$ between merged pairs) has controlled distortion under the Lipschitz assumption. It also provides a surrogate objective: penalise the expected state-space distance between collapsed pairs.

The implications are sharper than the familiar observation that lossy compression discards information. There exists a partial order on representations under which improving compression efficiency *monotonically* increases admissibility risk. Compression objectives unconstrained by admissibility exert structural pressure toward admissibility-violating identifications, strongest on rare high-consequence states—precisely the distinctions that appear as noise to the compressor but determine reachability at the boundaries that matter most.

3. Observational–Interventional Separation

Definition 4 (Observational Equivalence). $x_1 \sim_O x_2$ if $P(O_{t+1:t+k} \mid X_t = x_1) = P(O_{t+1:t+k} \mid X_t = x_2)$ for all $k \geq 1$ under a fixed observational policy.

Definition 5 (Interventional Equivalence). $x_1 \sim_I x_2$ if $\{y : x_1 \overset{\pi,h}{\rightsquigarrow} y\} = \{y : x_2 \overset{\pi,h}{\rightsquigarrow} y\}$ for all $\pi \in \Pi$, $h \leq H$.

Proposition 4. $\sim_I = \sim_A$.

Theorem 4 (Observational–Interventional Separation). *In general, $\sim_O \not\Rightarrow \sim_I$: there exist states $x_1, x_2 \in \mathcal{X}$ that are observationally equivalent but interventionally inequivalent.*

Canonical example. Let x_1 be a structurally intact load-bearing element and x_2 the same element with an internal crack invisible to surface sensors. Under passive observation, $P(O \mid x_1) = P(O \mid x_2)$, so $x_1 \sim_O x_2$. Under a policy applying stress beyond the crack threshold, x_2 admits failure states unreachable from x_1: $\mathcal{R}_H^{\Pi}(x_1) \neq \mathcal{R}_H^{\Pi}(x_2)$, so $x_1 \not\sim_I x_2$.

Remark 5 (Structural mismatch). The theorem identifies a *structural mismatch between equivalence-generating operations*: $\sim_O$ compares probability measures over passive observation sequences; $\sim_I$ compares subsets of $\mathcal{X}$ under arbitrary interventions. These are different mathematical objects. The gap cannot be closed by improving the predictive model, collecting more data, or shifting prediction from pixel space to latent space, because none of those operations changes the type of the equivalence relation being computed. The theorem targets *passive* representation learning; agents that act and observe outcomes generate interventional data and can in principle close the gap, but this requires active exploration, not self-supervised prediction.

The theorem is the planning-theoretic analogue of Pearl's do-calculus (Pearl, 2009): $P(Y \mid X = x) \neq P(Y \mid \mathrm{do}(X = x))$ in general; correspondingly, $\sim_O \neq \sim_I$ in general. It also formalises the identification–reachability gap in control theory (Ljung, 1999): system identification yields observational accuracy but not necessarily control adequacy. Both gaps are instances of a deeper inversion problem: the equivalence relation that passive observation generates is the wrong type for the question being asked. Standard frequentist methodology faces an exactly analogous inversion—it computes $P(\text{data} \mid \text{hypothesis})$ when the scientifically relevant quantity is $P(\text{hypothesis} \mid \text{data})$, and no accumulation of observational data resolves the type mismatch (Jaynes, 2003). The $\sim_O / \sim_A$ separation is the geometric form of this inversion: learners receive observations and compute observational partitions, but planning requires reachability partitions, and these are different mathematical objects.

A characterisation theorem (Appendix A) identifies precisely when coincidence holds: $\sim_O \subseteq \sim_A$ iff every intervention-relevant variable is observationally identifiable. This is an *additional* structural property of the system, not a consequence of predictive training. Verifying it is an independent problem that no current objective addresses.

4. Three Corollaries

4.1. Predictive Gap

Corollary 3 (Predictive Admissibility Gap)**.** *If φ_P collapses $x_1 \sim_O x_2$, then $D_A(\varphi_P) \not\equiv 0$ in general.*

By Theorem 4, there exist $x_1 \sim_O x_2$ with $\mathcal{R}^{\Pi}_H(x_1) \neq \mathcal{R}^{\Pi}_H(x_2)$. Any predictive projection identifying them contributes positively to $D_A(\varphi_P)$. This applies equally to contrastive predictive coding (van den Oord et al., 2018), successor representations (Dayan, 1993), and general value functions (Sutton et al., 2011): all encode future observation structure; all are subject to the predictive gap. Moving prediction from pixel space to latent space changes the domain of prediction without changing the type of the equivalence relation.

4.2. Compression Gap

Corollary 4 (Compression Admissibility Gap)**.** *$D_A(\varphi_C) \not\equiv 0$ in general for any projection φ_C minimising description length.*

By Theorem 3 and Corollary 1, each compression step incurs additional distortion equal to the expected separation of newly merged pairs. The information bottleneck (Tishby et al., 1999) specialises this: a representation that maximally compresses toward an observed reward signal may discard admissibility-relevant structure if that structure is uncorrelated with the reward in the training distribution.

4.3. Guardrail Incompleteness and Bayes-Risk Floor

Corollary 5 (Safety–Admissibility Gap)**.** *If $D_A(\varphi) > 0$ because $\varphi(x_s) = \varphi(x_d)$ for safe x_s and dangerous x_d, then no guardrail classifier operating on $\mathcal{M}$ alone (without temporal context, action history, or post-hoc outcome information) can distinguish them.*

Constructive proof. Any $G : \mathcal{M} \to \{0, 1\}$ satisfies $G(\varphi(x_s)) = G(\varphi(x_d))$ since both equal $G(z)$ for $z = \varphi(x_s) = \varphi(x_d)$. One state is necessarily misclassified. The failure is a consequence of the projection, not of classifier architecture or capacity. Monitors with access to action sequences or outcome histories may use information beyond $\varphi(x)$ and are not subject to this bound; the corollary applies to any classifier whose inputs are restricted to the latent representation at a single time step.

Theorem 5 (Irrecoverable Quotient Collapse)**.** *Let $A \in \{0, 1\}$ indicate admissibility status. For every downstream transformation $G : \mathcal{M} \to \{0, 1\}$,*

$$I(A;\, G(\varphi(X))) \;\leq\; I(A;\, \varphi(X))\,. \tag{2}$$

If $I(A; \varphi(X)) = 0$, then $I(A; G(\varphi(X))) = 0$ for all G.

(Proof in Appendix A.) This provides a principled diagnostic: persistent safety failure with error near 0.5 is attributable to the representation rather than the classifier architecture.

Corollary 6 (Bayes-Risk Floor)**.** *Let $Z = \varphi(X)$. The Bayes error of any binary classifier distinguishing $A \in \{0, 1\}$ from $\varphi(X)$ satisfies*

$$\varepsilon^* \;=\; \mathbb{E}_Z[\min\{P(A = 0 \mid Z),\; P(A = 1 \mid Z)\}]\,. \tag{3}$$

When $I(A; \varphi(X))$ is low, conditional uncertainty $H(A \mid Z)$ is high, and ε^* approaches $\min(p, 1 - p)$ where $p = P(A = 1)$. If $p \approx 0.5$, $\varepsilon^* \to 0.5$: any safety classifier on such a projection performs near-randomly regardless of its complexity.

5. Representation Learning as Quotient Selection

Proposition 5 (Quotient Framing)**.** *Every representation-learning method induces an equivalence relation $\sim_\varphi$ on $\mathcal{X}$ and thereby proposes a quotient $\mathcal{X}/\sim_\varphi$. The admissibility question is, in every case, whether $\sim_\varphi \subseteq \sim_A$.*

Anti-trivial-collapse mechanisms (EMA encoders, VICReg, contrastive terms) constrain the global distribution of equivalence classes without constraining which states are identified. They prevent $|\mathcal{X}/\sim_\varphi| = 1$ but do not enforce $\sim_\varphi \subseteq \sim_A$.

The disputes among JEPA, latent diffusion, transformer world models, bisimulation learning, and compression-based intelligence are, at this level of description, disputes about quotient selection. The renormalisation-group perspective (Wilson, 1975) asks which distinctions survive coarse-graining; admissibility inverts this: which distinctions *must not* be coarse-grained?

6. Admissibility as Prior Condition

The admissibility framework is not a competing objective: it is a prior condition derived from what planning requires, not assumed as a desideratum. Every planning goal—prediction, compression, control, reward, safety—presupposes that distinct futures can be distinguished and that some are accessible while others are not. Admissibility formalises that presupposition.

The claim is a necessary-condition claim, not a sufficiency claim. A system with $D_A(\varphi) \approx 0$ may still fail at understanding, consciousness, or adaptability for reasons this framework does not address. The framework claims only that $D_A(\varphi) > 0$ is sufficient for planning failure, and that $D_A(\varphi) = 0$ is necessary for adequate planning. This necessary condition is universal: it applies with equal force to biological nervous systems, social institutions, and artificial systems. A human planner who systematically confuses states with different reachable futures incurs planning failures by the same mechanism as a neural network with high $D_A(\varphi)$.

Model predictive control (Mayne et al., 2000; Garcia et al., 1989) presupposes that state variables are already correctly chosen—that the representation is a sufficient statistic for future evolution. The novel claim of JEPA-based world models is that self-supervised latent representations constitute such sufficient statistics. Corollary 3 establishes that this claim is unproven and that the criterion by which it could be evaluated is precisely $D_A(\varphi)$.

7. Toward Estimation

7.1. Foundational status

$D_A(\varphi)$ is not yet an engineering metric. It is a target criterion. The contribution of this paper is not an estimator; it is the identification of the quantity whose estimation future work must address. The research agenda is: develop estimators that are eventually tight in expectation, and certificates that are eventually sound.

7.2. Bisimulation lower bound

Proposition 6. *Let $D_{\text{bisim}}(\varphi)$ be the bisimulation distortion of φ under reward functions depending only on reachability. Then $D_{\text{bisim}}(\varphi) \le C \cdot D_A(\varphi)$ for a constant C depending on reward scale and discount factor. In particular, $D_A(\varphi) = 0$ implies $D_{\text{bisim}}(\varphi) = 0$.*

Bisimulation metrics (Ferns et al., 2004; Kemertas & Aumentado-Armstrong, 2021) thus provide computable lower bounds on $D_A(\varphi)$.

7.3. Trajectory-sampling lower bound

Fix $\Pi_k = \{\pi_1, \ldots, \pi_k\} \subseteq \Pi$. Define the *witnessed-disagreement indicator* for a collapsed pair as

$$W_k(x_1, x_2) = \mathbf{1}\Big[\exists\, \pi_i \in \Pi_k :\ \hat{\mathcal{R}}^{\pi_i, H}(x_1) \neq \hat{\mathcal{R}}^{\pi_i, H}(x_2)\Big],$$

where $\hat{\mathcal{R}}^{\pi_i,H}(x) = \{y : x \overset{\pi_i, H}{\rightsquigarrow} y\}$. Define the *witnessed distortion*

$$\hat{D}_A(\varphi; \Pi_k) = \mathbb{E}\big[\mathbf{1}_{C_\varphi = 1} \cdot W_k(x_1, x_2)\big].$$

Proposition 7 (Trajectory Lower Bound)**.** $\hat{D}_A(\varphi; \Pi_k) \le \mathbf{1}_{D_A(\varphi) > 0}$ *in the sense that $\hat{D}_A(\varphi; \Pi_k) > 0$ implies $D_A(\varphi) > 0$: any witnessed disagreement between collapsed states proves admissibility distortion exists. This bound does not require any metric monotonicity assumption.*

Proposition 8 (Asymmetry of Evidence)**.** *(i) If $\hat{D}_A(\varphi; \Pi_k) > 0$, then $D_A(\varphi) > 0$: any witnessed reachability disagreement between collapsed states* proves *distortion exists. (ii) If $\hat{D}_A(\varphi; \Pi_k) = 0$, this does not imply $D_A(\varphi) = 0$: a zero result proves only that the sampled policies found no witness.*

The asymmetry is that of model checking and adversarial testing: finding a counterexample is definitive; failing to find one is not a proof of correctness. Trajectory-sampling audits are one-sided diagnostic tools whose value increases with the adversarial diversity of Π_k.

7.4. Structural vs. operational distortion

Operational distortion $D_A^\pi(\varphi)$ is the expectation over states visited by the current policy; structural distortion $D_A^{\text{struct}}(\varphi) = \sup_{x_1, x_2} \mathbf{1}_{C_\varphi = 1} \cdot \Delta_A(x_1, x_2)$ measures the worst case over $\mathcal{X}$. These can diverge dramatically: $D_A^\pi \approx 0$ under a conservative policy while $D_A^{\text{struct}} \gg 0$ in rarely-visited regions. Safety-critical failures characteristically occur in low-frequency, high-consequence regions that operational distortion underweights.

7.5. Task-weighted distortion

A task-weighted refinement $D_A^w(\varphi) = \mathbb{E}[w(x_1, x_2) \cdot \mathbf{1}_{C_\varphi = 1} \cdot \Delta_A(x_1, x_2)]$ interpolates between operational and structural distortion when w assigns high weight to safety-critical regions regardless of visit frequency. The conflict between compression and admissibility is sharpest when the probability measure over states and the importance weights over states are systematically opposed, as they are in safety-critical domains.

7.6. Admissibility depth

Definition 6 (Admissibility Depth)**.** $d_A(x_1, x_2) = \min\{H : \mathcal{R}_H^\Pi[H](x_1) \neq \mathcal{R}_H^\Pi[H](x_2)\}$, with $d_A = \infty$ if $x_1 \sim_A x_2$ for all H.

The horizon hierarchy $\sim_A^{(\Pi,1)} \supseteq \sim_A^{(\Pi,10)} \supseteq \cdots$ implies that representations adequate for one-step prediction may be catastrophically inadmissible for strategic planning.

8. Discussion

Scope conditions. The framework presupposes a fixed $(\mathcal{X}, \Pi, H)$. When a system alters its own element types—new variables become relevant, old ones disappear—$\sim_A$ must be re-derived on the enlarged or altered space. This challenge is shared by every formal framework that requires a domain before analysis can begin. The admissibility framework is better positioned than distributional approaches: since $\sim_A$ is derived from planning requirements rather than training statistics, it does not presuppose a representative training distribution. Dynamic state-space evolution is an open problem for the field, not a special failure of admissibility-based approaches.

Relation to impossibility arguments. Arguments that intelligence is impossible to formalise because complex systems alter their own coordinate systems are addressed by the above scope observation. The admissibility relation is defined over reachability geometry, not distributional regularity; non-ergodicity does not preclude well-defined reachability cones. The relevant objection is not "irregularity implies impossible" but "changing coordinate systems complicate re-derivation of $\sim_A$"—a genuine difficulty that converts a putative refutation into an open research problem.

Self-modifying agents. An agent undergoing internal transformation (updated representations, goals, or memories) remains a coherent agent insofar as the transformation

preserves reachability-relevant distinctions. Narrative identity frameworks and symbolic stabilisers can be interpreted as mechanisms for maintaining low $D_A(\varphi)$ across developmental transitions, rather than alternatives to reachability-based accounts. Graziano's Attention Schema Theory (Graziano, 2013) provides a complementary angle: the model-of-the-world processor that tracks the agent's own attentional states is, in admissibility terms, tracking which of the agent's own reachable futures remain accessible given its current attention allocation. A low-distortion self-schema is a necessary condition for the attention schema to guide action reliably.

Relation to the Conscious Turing Machine. Blum & Blum (2022) define a formal seven-tuple model of consciousness in which a global broadcast mechanism (the Down Tree) simultaneously delivers a winning chunk to all long-term-memory processors. The chunk is selected by a probabilistic Up Tree tournament weighted by processor confidence. In admissibility terms, the broadcast mechanism is planning-adequate only if the chunk-selection process preserves $\sim_A$: two starting states whose reachable futures differ must not produce the same winning chunk. If the chunk representation (the *Brainish* gist) is inadmissible—if it collapses states with distinct reachability cones—then Corollary 5 applies: no downstream processor can recover the lost distinction, and the system's behaviour from those states will converge regardless of their different reachable futures. The CTM's subjective experience, grounded in the co-evolution of Brainish and a world model (Blum & Blum, 2022), is therefore admissibility-dependent: a genuine first-person perspective, in Varela et al.'s sense of "lived experience associated with cognitive and mental events" (Varela et al., 1999), requires that the internal language preserve the distinctions that determine what the agent can do next.

Relation to Gupta & Pruthi. Gupta & Pruthi (2025) argue that world models fail to capture understanding because state-tracking is insufficient for grasping the organising principles that make states significant. The admissibility framework formalises this diagnosis: the quantity world models fail to preserve is $D_A(\varphi)$, the departure of the learned quotient from $\mathcal{X}/\sim_A$. Poincaré's distinction between verifying a proof step-by-step and understanding why those steps were chosen (Poincaré, 1914) maps onto Theorem 4: verification is observational (local transition checking); understanding is interventional (grasping the reachability geometry of proof space).

Biological instantiation: photosynthesis–growth decoupling in oaks. A recent empirical study of North American temperate deciduous oaks provides what may be the clearest natural-system instantiation of the Observational–Interventional Separation Theorem outside of engineering (Rao et al., 2026). The authors demonstrate that gross primary productivity (GPP)—the observational quantity accessible to passive remote sensing, eddy covariance, and satellite fluorescence retrieval—and aboveground woody biomass growth—the quantity determining long-term carbon sequestration and thus the accessible futures of the ecosystem—are systematically decoupled across diel, seasonal, and interannual timescales. Across 137 tree-ring sites, 26–36% of annual GPP occurred after July, yet late-season climate contributed negligibly to annual ring-width variance. At four high-resolution monitoring sites, growth ceased 2–4 months before photosynthesis. At the daily scale, growth occurred primarily between midnight and 07:00 when VPD and temperature were minimal, while GPP peaked near solar noon under conditions of high atmospheric dryness.

The mechanistic explanation instantiates the coincidence-conditions theorem (Appendix A): observational equivalence implies interventional equivalence only when every intervention-relevant variable is observationally identifiable. Turgor pressure in cambial cells—the proximate driver of radial growth—is governed primarily by VPD and stem water status. GPP is governed primarily by light availability, stomatal conductance, and enzyme kinetics. These are different state variables subject to different biophysical constraints, and passive observation of carbon flux cannot identify which turgor state the cambium is in. Two oaks with identical GPP traces may therefore have radically different growth futures depending on their hydraulic status—exactly the structure of $x^h \sim_O x^f$ but $x^h \not\sim_A x^f$ in Theorem 4. Earth system models that conflate the photosynthetic season with the growing season are, in admissibility terms, using an inadmissible encoder: the learned quotient collapses states whose reachable futures—measured in woody carbon accumulated over decades—are substantially different.

The reported Spearman correlation $r = 0.86$ between interannual VPD variability and the degree of photosynthesis–growth decoupling (Rao et al., 2026) is an empirical measure of how admissibility distortion in the GPP representation scales with the activation of the latent constraint variable. As atmospheric aridity becomes more variable, the gap between observational and interventional equivalence widens—which is precisely what the framework predicts when a latent variable (turgor constraint) becomes more dynamically relevant. The monotonicity theorem (Theorem 3) further applies to the temporal aggregation practiced by ESMs: each step of coarsening from hourly to daily to seasonal to annual GPP is a compression step in the quotient lattice, and Corollary 1 identifies the admissibility cost as the expected reachability separation of the pairs each compression step newly merges. Rao et al. measure this cost directly: at annual resolution, post-July GPP carries near-zero information about annual ring-width, meaning the compression to annual totals has destroyed the phenological distinction between the growth window and the photosynthetic season entirely.

9. Conclusion

We have identified $D_A(\varphi)$ as the missing criterion in representation learning for planning, derived its properties from planning requirements alone, and established that predictive and compressive objectives fail to satisfy it in general. The Observational–Interventional Separation Theorem shows the failure is structural; the data-processing inequality and Fano bound convert it into operational safety limits; the decomposition and trajectory-sampling results provide diagnostic tools. $D_A(\varphi)$ joins the tradition of quantities—entropy, Kolmogorov complexity, mutual information, Wasserstein distance—that became organising centres for research programmes by naming something previously unmeasured. The contribution is not a new architecture or training procedure; it is the identification of a missing criterion and the derivation of its properties from first principles.

Acknowledgements. The author has no institutional affiliation and no conflicts of interest.

Appendix A. Proofs of All Results

We collect complete proofs of all propositions, theorems, and corollaries in the order of their appearance in the main text. Notation follows the main text throughout.

Proposition 1 (Planning Sufficiency Requirement)

Proof. Suppose $\mathcal{R}^{\Pi}_H(x_1) \neq \mathcal{R}^{\Pi}_H(x_2)$. Then there exists $y \in \mathcal{X}$ belonging to one reachability set but not the other; without loss of generality, $y \in \mathcal{R}^{\Pi}_H(x_1) \setminus \mathcal{R}^{\Pi}_H(x_2)$. Define a cost function $c : \mathcal{X} \to \mathbb{R}$ by $c(z) = 0$ if $z = y$ and $c(z) = M$ for $M > 0$ otherwise. Under any policy $\pi \in \Pi$, the expected cost from x_1 is achievable at zero (take π that reaches y); no policy from x_2 achieves cost less than $M > 0$ since $y \notin \mathcal{R}^{\Pi}_H(x_2)$. The optimal policies therefore differ. Any representation φ with $\varphi(x_1) = \varphi(x_2)$ cannot distinguish the starting state and hence cannot recover the optimal policy difference. □

Proposition 2 (Collapse–Separation Decomposition)

Proof. By the law of total expectation,

$$D_A(\varphi) = \mathbb{E}[C_\varphi \cdot \Delta_A] = \mathbb{E}[C_\varphi \cdot \Delta_A \mid C_\varphi = 1]\, P(C_\varphi = 1) + \mathbb{E}[C_\varphi \cdot \Delta_A \mid C_\varphi = 0]\, P(C_\varphi = 0).$$

The second term is zero since $C_\varphi = 0$ implies the indicator is zero. Hence $D_A(\varphi) = P(C_\varphi = 1) \cdot \mathbb{E}[\Delta_A \mid C_\varphi = 1]$. □

Theorem 2 (Universality of $\sim_A$)

Proof. ($\sim_A \subseteq \bigcap_{\mathcal{T}} \sim_{\mathcal{T}}$): If $x_1 \sim_A x_2$, then $\mathcal{R}^{\Pi}_H(x_1) = \mathcal{R}^{\Pi}_H(x_2)$. For any $\mathcal{T} \in \mathfrak{T}_A$ whose cost depends only on reachability, states with equal reachability sets induce identical optimal policies, so $x_1 \sim_{\mathcal{T}} x_2$.

($\bigcap_{\mathcal{T}} \sim_{\mathcal{T}} \subseteq \sim_A$): Suppose $\mathcal{R}^{\Pi}_H(x_1) \neq \mathcal{R}^{\Pi}_H(x_2)$. By Proposition 1, there is an accessibility task $\mathcal{T}^*$ (assign cost 0 to trajectories through the distinguishing reachable state, $M > 0$ otherwise) for which optimal policies differ, so $x_1 \not\sim_{\mathcal{T}^*} x_2$, hence $x_1 \notin \bigcap_{\mathcal{T}} \sim_{\mathcal{T}}$.

Maximality: any equivalence relation safe for all tasks in $\mathfrak{T}_A$ is contained in their intersection, which equals $\sim_A$. □

Theorem 3 (Monotonicity) and Corollary 1 (Accounting Identity)

Proof of Theorem 3. $\sim_{\varphi_1} \subseteq \sim_{\varphi_2}$ implies the expectation in $D_A(\varphi_2)$ sums over a superset of the pairs summed in $D_A(\varphi_1)$, with all terms non-negative. Hence $D_A(\varphi_1) \leq D_A(\varphi_2)$. □

Proof of Corollary 1. Since $\sim_{\varphi_2} = \sim_{\varphi_1} \cup \mathcal{N}$ (disjoint union by definition of $\mathcal{N}$),

$$D_A(\varphi_2) = \mathbb{E}[\mathbf{1}_{\sim_{\varphi_1}} \cdot \Delta_A] + \mathbb{E}[\mathbf{1}_{\mathcal{N}} \cdot \Delta_A] = D_A(\varphi_1) + \mathbb{E}[\mathbf{1}_{\mathcal{N}} \cdot \Delta_A].$$

□

Proposition 4 ($\sim_I = \sim_A$)

Proof. By definition, $x_1 \sim_I x_2$ iff their reachable sets coincide under all $\pi \in \Pi$ at all $h \leq H$. Taking $h = H$ and universally quantifying over π yields $\mathcal{R}^{\Pi}_H(x_1) = \mathcal{R}^{\Pi}_H(x_2)$, i.e. $x_1 \sim_A x_2$. The converse is immediate since $\mathcal{R}^{\Pi}_H(x) = \bigcup_{\pi \in \Pi} \{y : x \overset{\pi, H}{\rightsquigarrow} y\}$. □

Theorem 4 (Observational–Interventional Separation)

Proof. It suffices to exhibit a system in which $\sim_O \not\Rightarrow \sim_I$. The load-bearing element example serves: let x_1 (intact) and x_2 (internally cracked) satisfy $P(O_{t+1:t+k} \mid x_1) = P(O_{t+1:t+k} \mid x_2)$ for all k under passive observation (crack invisible at normal load), giving $x_1 \sim_O x_2$. Under a policy applying stress beyond the crack threshold, x_2 reaches failure states unreachable from x_1, so $\mathcal{R}^{\Pi}_H(x_1) \neq \mathcal{R}^{\Pi}_H(x_2)$ and $x_1 \not\sim_I x_2$.

A minimal formal example makes the structure explicit. Let $\mathcal{X} = \{s^0_0, s^1_0, s_{\text{safe}}, s_{\text{danger}}\}$ with latent binary variable $L \notin$ observation function. Set $O(s^0_0) = O(s^1_0) = o_0$, $O(s_{\text{safe}}) = o_s$, $O(s_{\text{danger}}) = o_d$. Under action a: $s^0_0 \to s_{\text{safe}}$; $s^1_0 \to s_{\text{danger}}$. Then $P(O_{t+1} \mid s^0_0) = P(O_{t+1} \mid s^1_0) = \delta_{o_0}$ so $s^0_0 \sim_O s^1_0$; but $\mathcal{R}^{\Pi}_H(s^0_0) = \{s_{\text{safe}}\}$, $\mathcal{R}^{\Pi}_H(s^1_0) = \{s_{\text{danger}}\}$, so $s^0_0 \not\sim_I s^1_0$.

More generally, any latent structural variable that determines reachability without affecting current observation distributions—a hidden Markov state, an unobservable causal parent, a dormant failure mode—produces the same pattern. Such variables are ubiquitous in physical, biological, and engineered systems. □

Observational Sufficiency Characterisation

Theorem 6 (Coincidence Conditions)**.** *$\sim_O \subseteq \sim_A$ iff every intervention-relevant variable is observationally identifiable: every variable whose value affects $\mathcal{R}^{\Pi}_H(x)$ is determined by the observation distribution $P(O_{t+1:t+k} \mid x)$.*

Proof. ($\Rightarrow$): If $\sim_O \subseteq \sim_A$ and some variable V affects reachability but is not observationally identified, there exist x_1, x_2 with $P(O \mid x_1) = P(O \mid x_2)$ but different V-values and different reachability sets, contradicting $\sim_O \subseteq \sim_A$.

($\Leftarrow$): If every intervention-relevant variable is observationally identifiable and $x_1 \sim_O x_2$, then every variable affecting $\mathcal{R}_H^\Pi(\cdot)$ takes the same value at x_1 and x_2, so $\mathcal{R}_H^\Pi(x_1) = \mathcal{R}_H^\Pi(x_2)$, giving $x_1 \sim_A x_2$. □

Theorem 5 (Irrecoverable Quotient Collapse)

Proof. The chain $A \to X \to \varphi(X) \to G(\varphi(X))$ satisfies the Markov property at each link: A causally determines X; φ is a deterministic function of X; G is a deterministic function of $\varphi(X)$. By the data-processing inequality (Cover & Thomas, 2006),

$$I(A;\, G(\varphi(X))) \leq I(A;\, \varphi(X)).$$

If $I(A;\varphi(X)) = 0$, then $I(A; G(\varphi(X))) \leq 0$; since mutual information is non-negative, equality holds. □

Theorem 1 (Zero-Distortion Characterisation)

Proof. $D_A(\varphi) = \mathbb{E}[C_\varphi \cdot \Delta_A]$. Every term is non-negative since $C_\varphi \geq 0$ and $\Delta_A = d(\mathcal{R}_H^\Pi(x_1), \mathcal{R}_H^\Pi(x_2)) \geq 0$. The expectation is zero iff every term is zero a.s. $C_\varphi(x_1, x_2) = 1$ and $\Delta_A > 0$ iff $\varphi(x_1) = \varphi(x_2)$ and $\mathcal{R}_H^\Pi(x_1) \neq \mathcal{R}_H^\Pi(x_2)$, i.e. $C_\varphi = 1$ and $x_1 \not\sim_A x_2$. Hence $D_A(\varphi) = 0$ iff $P(C_\varphi = 1, x_1 \not\sim_A x_2) = 0$. □

Corollary 6 (Bayes-Risk Floor)

Proof. For any classifier $G : \mathcal{M} \to \{0, 1\}$ and $Z = \varphi(X)$, the Bayes-optimal rule achieves the minimum posterior error at each z: $G^*(z) = \arg\max_{a \in \{0,1\}} P(A = a \mid Z = z)$. The resulting Bayes error is $\varepsilon^* = \mathbb{E}_Z[\min\{P(A = 0 \mid Z), P(A = 1 \mid Z)\}]$. By Theorem 5, no classifier on $\mathcal{M}$ can beat this floor. When $I(A; Z)$ is small, $P(A \mid Z) \approx P(A)$ for most z (Cover & Thomas, 2006), so $\min\{P(A = 0 \mid Z), P(A = 1 \mid Z)\} \approx \min(p, 1 - p)$ and $\varepsilon^* \approx \min(p, 1 - p)$. At $p = 0.5$, $\varepsilon^* \to 0.5$. □

Proposition 6 (Bisimulation Lower Bound)

Proof sketch. If $\mathcal{R}_H^\Pi(x_1) = \mathcal{R}_H^\Pi(x_2)$ (admissibility equivalence), states reach the same futures and have the same optimal value under any accessibility-structured reward, so bisimulation distance is zero. In the non-zero case, Lipschitz continuity of value functions with respect to reachability divergence under standard reward and discount assumptions (cf. Ferns et al. (2004)) gives $D_{\text{bisim}}(\varphi) \leq C \cdot D_A(\varphi)$ with C absorbing reward scale and $(1 - \gamma)^{-1}$. □

Proposition 7 (Trajectory Lower Bound)

Proof. If $\hat{D}_A(\varphi; \Pi_k) > 0$, there exists a collapsed pair (x_1, x_2) (with $\varphi(x_1) = \varphi(x_2)$) and a policy $\pi_i \in \Pi_k$ such that $\hat{\mathcal{R}}^{\pi_i, H}(x_1) \neq \hat{\mathcal{R}}^{\pi_i, H}(x_2)$. Since $\hat{\mathcal{R}}^{\pi_i, H}(x) \subseteq \mathcal{R}_H^\Pi(x)$, we have $\mathcal{R}_H^\Pi(x_1) \neq \mathcal{R}_H^\Pi(x_2)$ (otherwise both empirical sets would be subsets of the same true set and the collapsed pair could not exhibit disagreement). Hence $x_1 \not\sim_A x_2$, giving $D_A(\varphi) > 0$ by Theorem 1. The bound requires no metric monotonicity assumption. □

Proposition 8 (Asymmetry of Evidence)

Proof. Part (i): $\hat{D}_A > 0$ implies $D_A(\varphi) \geq \hat{D}_A > 0$ by Proposition 7.

Part (ii): $\hat{D}_A = 0$ means that for every collapsed pair (x_1, x_2) in the sample, the sampled trajectories from x_1 and x_2 under Π_k did not diverge within horizon H. Since $\hat{\mathcal{R}}^{\Pi_k, H}(x) \subseteq \mathcal{R}_H^\Pi(x)$ with potentially strict inclusion (when $\Pi_k \subsetneq \Pi$ or trajectories are censored), non-divergence under Π_k does not imply $\mathcal{R}_H^\Pi(x_1) = \mathcal{R}_H^\Pi(x_2)$. Hence $\hat{D}_A = 0$ does not imply $D_A(\varphi) = 0$. □

Appendix B. Stochastic Extension

The main text treats deterministic dynamics for clarity. Replace $\mathcal{R}_H^\Pi(x)$ by the distribution $\mathcal{P}(\cdot \mid x, \pi, H)$ over future trajectories. Admissibility equivalence becomes equality of these distributions for all $\pi \in \Pi$, and admissibility distortion becomes

$$D_A(\varphi)^{\text{stoch}} = \mathbb{E}_{x_1, x_2}\left[\mathbf{1}_{\varphi(x_1) = \varphi(x_2)} \cdot S_p(x_1, x_2)\right], \quad (4)$$

where $S_p(x_1, x_2) = \sup_{\pi \in \Pi} W_p(\mathcal{P}(\cdot \mid x_1, \pi, H), \mathcal{P}(\cdot \mid x_2, \pi, H))$ where W_p is the Wasserstein-p distance (KL divergence is an alternative with tighter information-theoretic connections). All main theorems carry over: the Separation Theorem holds because passive observation and intervention still generate distributions over different mathematical objects; the data-processing inequality applies without modification; the bisimulation lower bound extends via the stochastic bisimulation metric of Ferns et al. (2004); and the trajectory lower bound extends by replacing set inclusion with distributional domination under the chosen divergence.

Appendix C. Admissibility Depth and Horizon Hierarchy

Admissibility equivalence is parameterised by (Π, H). Write $\sim_A^{(\Pi,H)}$ explicitly. A natural hierarchy holds:

$$\sim_A^{(\Pi,1)} \supseteq \sim_A^{(\Pi,10)} \supseteq \cdots \supseteq \sim_A^{(\Pi,\infty)}.$$

Definition 7 (Admissibility Depth)**.** $d_A(x_1, x_2) = \min\{H \in \mathbb{N} : x_1 \not\sim_A^{(\Pi,H)} x_2\}$, with $d_A(x_1, x_2) = \infty$ if $x_1 \sim_A^{(\Pi,H)} x_2$ for all H.

Large admissibility depth means two states appear equivalent for short-horizon planning but diverge at longer horizons. A representation adequate for one-step prediction ($H = 1$) may be catastrophically inadmissible for strategic planning ($H \gg 1$), even when $D_A^{(\Pi,1)}(\varphi) = 0$.

Proposition 9 (Distortion Monotonicity in Horizon)**.** *For fixed* (Π, φ)*,* $D_A^{(\Pi,H)}(\varphi)$ *is non-decreasing in* H*.*

Proof. Since Definition 1 defines reachability as "*within* horizon H" (not exactly at H), we have $\mathcal{R}_H^\Pi[H](x) \subseteq \mathcal{R}_H^\Pi[H+1](x)$ for all x and H: any state reachable within H steps is also reachable within $H+1$ steps (by executing the same policy and then idling or repeating the last action). Therefore $\sim_A^{(\Pi,H+1)} \subseteq \sim_A^{(\Pi,H)}$: longer horizons can distinguish states that shorter horizons cannot. By Theorem 3 applied to the partition induced by φ relative to the changing $\sim_A$, distortion is non-decreasing in H. □

Proposition 3 (Lipschitz Stability)

Proof. By assumption, $\Delta_A(x_1, x_2) \leq L\, d_{\mathcal{X}}(x_1, x_2)$ for all x_1, x_2. Substituting into Definition 3:

$$\begin{aligned} D_A(\varphi) &= \mathbb{E}[C_\varphi \cdot \Delta_A] \leq L\, \mathbb{E}[C_\varphi \cdot d_{\mathcal{X}}(x_1, x_2)] \\ &= L\, \mathbb{E}[\mathbf{1}_{C_\varphi = 1} \cdot d_{\mathcal{X}}(x_1, x_2)]. \end{aligned}$$ □

Appendix D. Admissibility-Constrained Information Bottleneck

The information bottleneck (Tishby et al., 1999) seeks a representation Z of X minimising $I(X;Z)$ subject to $I(Z;Y) \geq \beta$ for a target variable Y and tradeoff $\beta > 0$. We propose the *admissibility-constrained bottleneck*: minimise $I(X;Z)$ subject to $D_A(\varphi) \leq \epsilon$ for some tolerance $\epsilon \geq 0$.

The standard bottleneck can discard admissibility-relevant structure if it is uncorrelated with Y in the training distribution. The admissibility-constrained variant makes this trade-off explicit: description length is minimised only among projections that preserve reachability-relevant distinctions to within ϵ. At $\epsilon = 0$, the feasible set is exactly the set of admissible projections; at $\epsilon > 0$, controlled inadmissibility is permitted in exchange for compression efficiency.

The dual formulation adds $\lambda \cdot D_A(\varphi)$ to the bottleneck objective:

$$\min_\varphi \; \mathcal{L}(\varphi, \lambda) = I(X; \varphi(X)) + \lambda \cdot D_A(\varphi).$$

Proposition 10 (Lagrangian Limiting Cases)**.** *Let* φ_λ *denote the minimiser of* $\mathcal{L}(\cdot, \lambda)$*.*

(i) $\lambda = 0$*: reduces to pure compression;* $D_A(\varphi)$ *is unconstrained.*

(ii) $\lambda \to \infty$*:* $D_A(\varphi_\lambda) \to 0$*; the solution approaches the minimally compressive admissible projection.*

(iii) Under standard envelope-theorem conditions on the parametric family of minimisers,

$$\frac{d}{d\lambda}\, \mathcal{L}(\varphi_\lambda, \lambda) = D_A(\varphi_\lambda).$$

Proof sketch. (i) and (ii) follow immediately from the objective: at $\lambda = 0$ the admissibility term vanishes; as $\lambda \to \infty$ any φ with $D_A(\varphi) > 0$ has unbounded cost and is dominated by any admissible projection. (iii) is the Danskin–Rockafellar envelope theorem: $\partial\mathcal{L}/\partial\lambda$ evaluated at φ_λ equals the partial derivative in λ with φ held fixed, which is $D_A(\varphi)(\varphi_\lambda)$. □

The envelope result (iii) implies that the rate of change of the optimal objective with respect to the admissibility penalty equals the current distortion level. A steep rate of change at small λ indicates that compression is being achieved primarily by sacrificing admissibility; a flat rate indicates that admissibility and compression are compatible in the current representation family. Developing tractable surrogates for $D_A(\varphi)$ in this objective—via bisimulation distances, trajectory-divergence penalties, or reachability-coverage regularisers—is a concrete algorithmic research problem.

The admissibility penalty $\lambda \cdot D_A(\varphi)$ plays a role structurally dual to the Bayesian Occam's razor (Jaynes, 2003). The Bayesian Bayes factor penalises models for occupying unnecessary parameter dimensions: prior probability density becomes diluted across a larger space, dragging down the marginal likelihood unless the extra complexity yields compensating predictive accuracy. The admissibility penalty penalises representations for collapsing necessary distinctions: $D_A(\varphi)$ accumulates whenever the projection merges states whose reachable futures differ, and this cost grows monotonically with compression (Theorem 3). The two mechanisms are mirror images — one punishes unnecessary complexity; the other punishes destructive simplicity — and together they define the boundaries of a well-posed representation space for planning.

Appendix E. Dynamic State Spaces and Meta-Admissibility

The main framework takes $(\mathcal{X}, \Pi, H)$ as fixed. When a system alters its own element types—new state dimensions become relevant, old ones disappear—$\sim_A$ must be re-derived on the evolved space. Let $\mathcal{X}_t$ denote the state space at time t and $\sim_A^{(t)}$ the corresponding admissibility relation.

A transformation $F : \mathcal{X}_t \to \mathcal{X}_{t+1}$ (representing a developmental or ontological change in the system) is *admissibility-preserving* if the induced map on equivalence classes satisfies:

$$x_1 \sim_A^{(t)} x_2 \;\Rightarrow\; F(x_1) \sim_A^{(t+1)} F(x_2).$$

This is a necessary condition for the transformation to preserve planning competence across the transition.

Definition 8 (Meta-Admissibility)**.** A sequence of transitions $(\mathcal{X}_t, \sim_A^{(t)}) \to (\mathcal{X}_{t+1}, \sim_A^{(t+1)})$ is *meta-admissible* if the transition map F is admissibility-preserving.

The dynamic vector-space objection (complex systems alter their own element types, making fixed mathematical structures inapplicable) is therefore not a refutation

of admissibility theory but a request for its generalisation to evolving state spaces. The question shifts from "does a projection preserve reachability?" to "does a state-space evolution preserve reachability?" These are formally analogous questions one level up. Narrative identity frameworks, symbolic stabilisers, and developmental coherence conditions can be interpreted as mechanisms for enforcing meta-admissibility: they constrain which self-modifications are permissible by requiring that admissibility-relevant distinctions survive transition.

Meta-admissibility is an open research programme. The present paper establishes the base case (t fixed) and identifies the natural generalisation.

Proposition 11 (Composition of Admissibility-Preserving Maps)**.** *If $F : \mathcal{X}_t \to \mathcal{X}_{t+1}$ and $G : \mathcal{X}_{t+1} \to \mathcal{X}_{t+2}$ are both admissibility-preserving, then $G \circ F : \mathcal{X}_t \to \mathcal{X}_{t+2}$ is admissibility-preserving.*

Proof. Let $x_1 \sim_A^{(t)} x_2$. Since F is admissibility-preserving, $F(x_1) \sim_A^{(t+1)} F(x_2)$. Since G is admissibility-preserving, $G(F(x_1)) \sim_A^{(t+2)} G(F(x_2))$. □

Admissibility-preserving maps therefore compose, giving the collection of admissibility-preserving transitions a category-like structure. The objects are pairs $(\mathcal{X}_t, \sim_A^{(t)})$ and the morphisms are admissibility-preserving maps. This means sequences of developmental transitions can be analysed by decomposing them into individual admissibility-preserving steps: if each step is admissibility-preserving, the composition over any finite developmental trajectory is admissibility-preserving. Conversely, a single non-admissibility-preserving transition contaminates all subsequent states reachable from it.

Appendix F. Relation to Bisimulation Hierarchy

Bisimulation equivalence $\sim_{\text{bis}}$ requires identical transition distributions and reward structures: $x_1 \sim_{\text{bis}} x_2$ iff $P(\cdot \mid x_1, a) = P(\cdot \mid x_2, a)$ and $r(x_1, a) = r(x_2, a)$ for all a. Admissibility equivalence is strictly coarser: it requires only that reachable future *sets* coincide, not that transition distributions or rewards agree pointwise.

Proposition 12. *$\sim_{\text{bis}} \subseteq \sim_A$ for reward functions depending only on reachability structure.*

Proof. If $x_1 \sim_{\text{bis}} x_2$, then transition distributions are identical under all actions; by induction over H, reachable sets are identical, so $\mathcal{R}_H^{\Pi}(x_1) = \mathcal{R}_H^{\Pi}(x_2)$, giving $x_1 \sim_A x_2$. □

The converse fails in general: two states can have $\mathcal{R}_H^{\Pi}(x_1) = \mathcal{R}_H^{\Pi}(x_2)$ (and hence $x_1 \sim_A x_2$) while differing in transition dynamics or reward (failing $\sim_{\text{bis}}$). Admissibility is therefore a *weaker* requirement than bisimulation: it demands only that planning-relevant structure be preserved, not that the full stochastic dynamics be matched.

This hierarchy has a practical implication. Bisimulation-based representation learning (Zhang et al., 2021; Kemertas & Aumentado-Armstrong, 2021) optimises for a more stringent criterion than admissibility requires. Systems satisfying bisimulation equivalence are admissible, but admissibility does not require bisimulation. A representation with $D_A(\varphi) = 0$ but $D_{\text{bisim}} > 0$ is adequate for planning purposes despite failing the bisimulation criterion. Conversely, Proposition 6 shows that D_{bisim} lower-bounds a rescaling of $D_A(\varphi)$, so bisimulation distortion serves as a useful computable proxy.

Appendix G. Category of Admissibility Structures

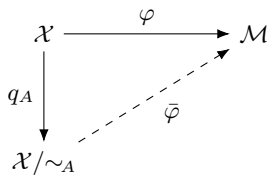
We show that admissibility structures form a category in which admissible projections are precisely the morphisms, and that the admissibility quotient $q_A : \mathcal{X} \to \mathcal{X}/\!\sim_A$ satisfies a universal property making it initial among all planning-safe abstractions.

Definition 9 (Category **Adm**)**.** The category **Adm** has:

- *Objects*: pairs $(\mathcal{X}, \sim_A)$ where $\mathcal{X}$ is a state space and $\sim_A$ is an admissibility equivalence relation on $\mathcal{X}$.
- *Morphisms*: functions $f : (\mathcal{X}, \sim_A) \to (\mathcal{Y}, \sim_A')$ satisfying $x_1 \sim_A x_2 \Rightarrow f(x_1) \sim_A' f(x_2)$.
- *Composition*: function composition (well-defined since admissibility-preservation is preserved under composition, Proposition 11).
- *Identity*: $\mathrm{id}_{\mathcal{X}} : (\mathcal{X}, \sim_A) \to (\mathcal{X}, \sim_A)$.

Admissible projections $\varphi : \mathcal{X} \to \mathcal{M}$ are exactly the morphisms $(\mathcal{X}, \sim_A) \to (\mathcal{M}, \sim_A')$ in **Adm** when $\mathcal{M}$ is equipped with the finest admissibility structure consistent with φ. Inadmissible projections fail to be morphisms: there exist $x_1 \sim_A x_2$ with $\varphi(x_1) \neq \varphi(x_2)$ in $\mathcal{M}$ where these images are not equivalent. Admissibility distortion $D_A(\varphi)$ measures the extent of this morphism failure.

Theorem 7 (Universal Property of q_A)**.** *Let $q_A : \mathcal{X} \to \mathcal{X}/\!\sim_A$ be the canonical quotient map. For any admissible projection $\varphi : (\mathcal{X}, \sim_A) \to (\mathcal{M}, \sim_A')$ in* **Adm***, there exists a unique morphism $\bar{\varphi} : (\mathcal{X}/\!\sim_A, \sim_A^*) \to (\mathcal{M}, \sim_A')$ such that $\varphi = \bar{\varphi} \circ q_A$:*



(5)

where $\sim_A^*$ is the quotient relation on $\mathcal{X}/\!\sim_A$ induced by $\sim_A$. Consequently, q_A is the initial object *in the category of admissible projections from $\mathcal{X}$: every planning-safe representation factors uniquely through the admissibility quotient.*

Proof. Existence. Define $\bar{\varphi} : [x] \mapsto \varphi(x)$ for each equivalence class $[x] \in \mathcal{X}/\!\sim_A$. Since φ is admissible, if

$x_1 \sim_A x_2$ then $\varphi(x_1) = \varphi(x_2)$ (by Definition 2), so $\bar{\varphi}$ is well-defined on equivalence classes. By construction, $\bar{\varphi}(q_A(x)) = \bar{\varphi}([x]) = \varphi(x)$, so $\varphi = \bar{\varphi} \circ q_A$.

Uniqueness. Any $\psi : \mathcal{X}/\sim_A \to \mathcal{M}$ satisfying $\varphi = \psi \circ q_A$ must satisfy $\psi([x]) = \psi(q_A(x)) = \varphi(x)$ for all x. Since every element of $\mathcal{X}/\sim_A$ is of the form $[x]$, ψ is uniquely determined.

Initiality. A morphism $q_A : (\mathcal{X}, \sim_A) \to (\mathcal{X}/\sim_A, \sim_A^*)$ exists and is universal: every morphism out of $(\mathcal{X}, \sim_A)$ factors uniquely through it. This is the defining property of an initial object in the slice category over $\mathcal{M}$. □

Theorem 7 gives the admissibility quotient a character stronger than "largest safe equivalence relation": it is the *initial* planning-safe abstraction, the one through which every other safe abstraction necessarily factors. The factorisation map $\bar{\varphi}$ is unique, making the relationship between any admissible representation and the admissibility quotient canonical rather than constructed.

The categorical language also reframes admissibility distortion. $D_A(\varphi)$ measures how far φ is from being a morphism in **Adm**: specifically, it measures the expected reachability separation between pairs that φ collapses but $\sim_A$ does not. A zero-distortion projection is a morphism; a positive-distortion projection is a broken morphism whose degree of failure is quantified by $D_A(\varphi)$.

The lattice structure of **Adm** connects to the Knuth–Skilling programme (Knuth & Skilling, 2012), which derives the sum and product rules of probability as the unique consistent valuation on a distributive lattice ordered by logical implication. The admissibility quotient lattice $(\mathcal{X}/\sim_A, \subseteq)$, ordered by set inclusion of reachability sets, is a distributive lattice of exactly this type. A valuation on this lattice that assigns 0 to admissibility-equivalent pairs and positive values to inadmissible collapses is precisely $D_A(\varphi)$. The Knuth–Skilling uniqueness result suggests that $D_A(\varphi)$, up to choice of severity metric d, is the canonical measure of departure from morphism-hood in **Adm**—the unique consistent quantification of admissibility violation, just as probability is the unique consistent quantification of logical implication on a distributive lattice of propositions.

Appendix H. Sheaf-Theoretic Reading of the Separation Theorem

We show that the Observational–Interventional Separation Theorem (Theorem 4) has a natural sheaf-theoretic interpretation: observation provides sections of one sheaf while planning requires sections of a different sheaf on a finer topology. The impossibility of recovering planning-relevant information from observational data corresponds to a sheaf cohomological obstruction.

Reachability Topology and the Planning Sheaf

Define the *reachability topology* τ_A on $\mathcal{X}$ as the topology generated by the reachability balls

$$B_H^\Pi(x) = \mathcal{R}_H^\Pi(x) = \{y \in \mathcal{X} : \exists \pi \in \Pi,\ x \overset{\pi,H}{\rightsquigarrow} y\},$$

taking these as basic open sets. The topology τ_A captures which states are "near" x in the sense of being reachable from it.

Define the *planning presheaf* $\mathcal{F}_A$ by assigning to each open $U \in \tau_A$ the set of reachable futures from U:

$$\mathcal{F}_A(U) = \bigcup_{x \in U} \mathcal{R}_H^\Pi(x),$$

with restriction maps $\rho_{UV} : \mathcal{F}_A(U) \to \mathcal{F}_A(V)$ given by inclusion for $V \subseteq U$.

Observation Topology and the Observation Sheaf

Define the *observation topology* τ_O on $\mathcal{X}$ as generated by observational indistinguishability neighborhoods:

$$N_O(x) = \big\{y \in \mathcal{X} : P(O_{t+1:t+k} \mid y) = P(O_{t+1:t+k} \mid x) \text{ for all } k \geq 1\big\}.$$

The observation sheaf $\mathcal{F}_O$ assigns to $U \in \tau_O$ the observation distributions accessible from U:

$$\mathcal{F}_O(U) = \big\{P(O_{t+1:t+k} \mid x) : x \in U,\ k \geq 1\big\}.$$

The Separation Theorem as a Sheaf Statement

Theorem 8 (Sheaf Formulation of Separation)**.** *The topologies τ_O and τ_A are distinct in general, with τ_O coarser than τ_A: observationally indistinguishable states may be reachability-distinguishable. Consequently, a global section of $\mathcal{F}_A$ (complete reachability information for all of $\mathcal{X}$) is not determined by the sections of $\mathcal{F}_O$ (observational information on a τ_O-cover). There exist covers $\{U_i\}$ of $\mathcal{X}$ in τ_O on which local sections of $\mathcal{F}_O$ are compatible, yet do not assemble into a section of $\mathcal{F}_A$.*

Proof. By Theorem 4, there exist $x_1, x_2 \in \mathcal{X}$ with $x_1 \sim_O x_2$ (so x_1 and x_2 belong to the same τ_O-neighborhood) but $\mathcal{R}_H^\Pi(x_1) \neq \mathcal{R}_H^\Pi(x_2)$ (so x_1 and x_2 belong to different τ_A-basic-open-sets). Any τ_O-cover of $\mathcal{X}$ groups x_1 and x_2 into the same open set and assigns them the same local section of $\mathcal{F}_O$. A global section of $\mathcal{F}_A$ must distinguish x_1 and x_2 (since their reachability sets differ), but no τ_O-local information can force this distinction. The attempted assembly of local $\mathcal{F}_O$-sections into a global $\mathcal{F}_A$-section therefore fails the uniqueness condition of the sheaf axiom: two distinct global admissibility sections are consistent with the same observational local data. □

Remark 6 (Interpretation). Local sections of $\mathcal{F}_O$ correspond to passive observational data available to a predictive

learner. A global section of $\mathcal{F}_A$ corresponds to complete knowledge of the reachability geometry required for planning. Theorem 8 establishes that observation provides local sections of $\mathcal{F}_O$, while planning requires a global section of $\mathcal{F}_A$, and these are sections of different sheaves on different topologies. The failure of local-to-global assembly is precisely the observational–interventional gap: locally consistent observation does not determine globally consistent reachability.

The failure of assembly can be understood as a cohomological obstruction: the mismatch between τ_O and τ_A produces a non-trivial Čech cohomology class $[\delta s] \in H^1(\tau_O; \mathcal{F}_A)$ that obstructs extension of local $\mathcal{F}_O$-sections to global $\mathcal{F}_A$-sections. Formalising this obstruction class and connecting it to the mutual-information bound $I(A; \varphi(X))$ of Theorem 5 is a natural direction for future work: the magnitude of the obstruction may be quantifiable in terms of $D_A(\varphi)$.

Temporal Admissibility as a Sheaf over Time

The dynamic state-space setting of Appendix E also admits a sheaf-theoretic reading. Let the base category be the poset of time intervals $[s, t]$ ordered by inclusion. Assign to each interval $[s, t]$ the admissibility structure $(\mathcal{X}_{[s,t]}, \sim_A^{[s,t]})$ on the time-slice of state space. The restriction maps are the transition functions $F : \mathcal{X}_t \to \mathcal{X}_{t+1}$ of Appendix E. Meta-admissibility (Definition in Appendix E) is then precisely the sheaf compatibility condition: sections on overlapping time intervals agree on their overlap. The composition theorem (Proposition 11) is the sheaf gluing lemma for this temporal sheaf: compatible local admissibility structures assemble into a global admissibility structure over the developmental trajectory. A developmental transition that violates meta-admissibility corresponds to a failed gluing, creating a cohomological obstruction to consistent identity over time.

References

Bertsekas, D. P. (2012). *Dynamic Programming and Optimal Control*, Vol. 1, 4th ed. Athena Scientific.

Blum, L., & Blum, M. (2022). A theoretical computer science perspective on consciousness. *Journal of Artificial Intelligence and Consciousness*, 9(1), 1–42.

Castro, P. S., & Precup, D. (2010). Using bisimulation for policy transfer in MDPs. *AAAI*.

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley.

Craik, K. J. W. (1943). *The Nature of Explanation*. Cambridge University Press.

Dayan, P. (1993). Improving generalisation for temporal difference learning: The successor representation. *Neural Computation*, 5(4), 613–624.

Ferns, N., Panangaden, P., & Precup, D. (2004). Metrics for finite Markov decision processes. *UAI*, 162–169.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. A*, 222, 309–368.

Garcia, C. E., Prett, D. M., & Morari, M. (1989). Model predictive control: Theory and practice. *Automatica*, 25(3), 335–348.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Graziano, M. S. A. (2013). *Consciousness and the Social Brain*. Oxford University Press.

Rao, M. P., Pacheco-Solana, A., Li, R., et al. (2026). Decoupled carbon assimilation and growth responses to aridity in temperate deciduous oaks. *Science Advances*, 12, eady7139.

Gupta, T., & Pruthi, D. (2025). Beyond world models: Rethinking understanding in AI models. *arXiv*:2511.12239.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Knuth, K. H., & Skilling, J. (2012). Foundations of inference. *Axioms*, 1(1), 38–73.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2), 99–134.

Kalman, R. E. (1960). On the general theory of control systems. *Proc. First IFAC Congress*, 481–492.

Kemertas, M., & Aumentado-Armstrong, T. (2021). Towards robust bisimulation metric learning. *NeurIPS*, 34.

Li, L., Walsh, T. J., & Littman, M. L. (2006). Towards a unified theory of state abstraction for MDPs. *ISAIM*, 531–539.

Littman, M. L., Sutton, R. S., & Singh, S. (2002). Predictive representations of state. *NeurIPS*, 14.

Ljung, L. (1999). *System Identification: Theory for the User*, 2nd ed. Prentice Hall.

Mayne, D. Q., Rawlings, J. B., Rao, C. V., & Scokaert, P. O. M. (2000). Constrained model predictive control: Stability and optimality. *Automatica*, 36(6), 789–814.

van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv*:1807.03748.

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

Poincaré, H. (1914). *Science and Method*. Thomas Nelson.

Popper, K. R. (1979). *Objective Knowledge*. Clarendon Press.

Ravindran, B., & Barto, A. G. (2004). Approximate homomorphisms: A framework for non-exact minimization in MDPs. *KBCS*.

Schmidhuber, J. (1991). A possibility for implementing curiosity and boredom. *SAB*, 222–227.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation. *IEEE Trans. Auton. Mental Dev.*, 2(3), 230–247.

Sutton, R. S., et al. (2011). Horde: A scalable real-time architecture for learning knowledge. *AAMAS*, 761–768.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. *Allerton*, 368–377.

Tomlin, C. J., Mitchell, I., Bayen, A. M., & Oishi, M. (2003). Computational techniques for verification of hybrid systems. *Proc. IEEE*, 91(7), 986–1001.

Varela, F. J., Shear, J., *et al.* (Eds.) (1999). *The View from Within: First-Person Approaches to the Study of Consciousness*. Imprint Academic.

Véliz, C. (2024). *Prophecy: The Power and Perils of Looking into the Future*. Oxford Univ. Press.

Wilson, K. G. (1975). The renormalization group: Critical phenomena and the Kondo problem. *Rev. Mod. Phys.*, 47(4), 773–840.

Zhang, A., McAllister, R., Calandra, R., Gal, Y., & Levine, S. (2021). Learning invariant representations for RL without reconstruction. *ICLR*.