Admissibility Distortion in Latent State Encoders: A Reachability-Geometric Framework for Fault-Critical Mechatronic Systems

Flyxion
Independent Researcher

June 2026

Abstract

State encoders embedded in model-based controllers for mechatronic systems are conventionally evaluated by predictive fidelity metrics—reconstruction error, prediction horizon accuracy, or spectral coherence between encoder output and measured plant response. We demonstrate that these metrics are insufficient for fault-critical planning and control: a formally stated *Observational–Interventional Separation Theorem* establishes that encoder observational equivalence does not imply interventional equivalence, and identifies this as a structural property of the encoder's induced quotient map rather than an empirical deficiency correctable by additional training data. The quantity that predictive metrics fail to bound is *admissibility distortion* $D_A(\boldsymbol{\varphi})$, defined as the expected Hausdorff distance between reachability sets of encoder-collapsed state pairs. We derive its decomposition into collapse-rate and per-collapse severity factors, prove strict monotonicity under compression, and establish a Bayes-risk floor on any downstream fault classifier operating on an inadmissible encoder output. Trajectory-rollout auditing procedures analogous to hardware-in-the-loop (HIL) fault injection are derived as one-sided witnesses for $D_A(\boldsymbol{\varphi}) > 0$. Applications to brushless DC motor drive fault isolation, structural health monitoring (SHM) with piezoelectric sensor arrays, and inverter-fed induction motor flux observers are developed. The framework subsumes bisimulation-based abstractions and provides a computable lower bound via bisimulation distortion metrics already in use for model-based RL controllers.

Index Terms: admissibility distortion, reachability analysis, state encoder, fault isolation, mechatronics, brushless DC motor, structural health monitoring, flux observer, model-based control, HIL auditing.

1. Introduction

Modern mechatronic systems rely on latent-state encoders—neural networks, Kalman filter banks, principal-component projections, or learned embedding networks—to compress high-dimensional sensor data into compact state vectors suitable for model predictive controllers (MPC) (Mayne et al., 2000), reinforcement learning (RL) policy networks (Bertsekas, 2012), or fault-diagnostic pipelines. The standard design criterion for such encoders is predictive accuracy: the encoder $\boldsymbol{\varphi} : \mathcal{X} \to \mathcal{M}$ is accepted if it minimises multi-step prediction error on held-out plant trajectories.

This paper establishes that predictive accuracy is *insufficient* for fault-critical applications. The argument proceeds in three stages. First, we define the *admissibility equivalence* $\sim_A$ as the coarsest equivalence on plant state space $\mathcal{X}$ compatible with planning-adequate discrimination. Second, we prove that observational equivalence $\sim_O$ (the equivalence induced by predictive training) does not imply $\sim_A$ in general—a result we call the Observational–Interventional Separation Theorem (OIST). Third, we quantify the gap as *admissibility distortion* $D_A(\boldsymbol{\varphi})$ and derive its engineering consequences: a monotone compression penalty, a Bayes-risk floor on fault classifiers, and an HIL-compatible auditing procedure.

1.1. Motivating scenario: BLDC motor drive

Consider a three-phase brushless DC (BLDC) motor drive (Fig. 1) controlled by a field-oriented controller (FOC) with a learned encoder compressing phase-current measurements $\{i_a, i_b, i_c\}$ and rotor position θ_r into a latent vector $\boldsymbol{z} \in \mathbb{R}^d$. Under healthy operation and incipient inter-turn short-circuit (ITSC) fault, the Clarke-transformed currents i_α, i_β may exhibit nearly identical spectral content below the fundamental harmonic—the faults are observationally equivalent under passive sinusoidal excitation. Under a field-weakening intervention (injection of a high-frequency probing signal v_{hf} as in Holtz (2006)), the two conditions produce divergent current trajectories because the ITSC fault alters the effective d-axis inductance L_d, which determines the reachable trajectory set under the probing intervention. An encoder trained exclusively on steady-state prediction merges the two conditions; a fault classifier operating on its output is then guaranteed to perform near-randomly on ITSC detection—not due to classifier inadequacy, but because the encoder has destroyed the relevant distinction.

1.2. Paper organisation

Section II defines the plant model and reachability formalism. Section III derives admissibility distortion and its

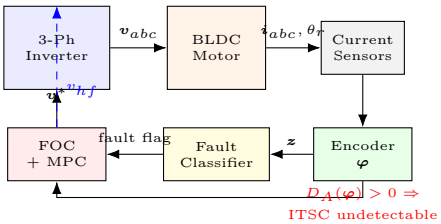

Figure 1: BLDC FOC drive with learned encoder φ and fault classifier. Dashed line: high-frequency probing injection for ITSC detection. If the encoder collapses the healthy and ITSC states (identical passive observations), the fault classifier operates on a zero-information projection.

decomposition. Section IV states and proves the OIST. Section V derives engineering corollaries for fault classification and safety. Section VI develops the HIL auditing procedure. Section VII presents three application case studies. Section VIII concludes.

2. Plant Model and Reachability Geometry

2.1. General mechatronic state-space model

Let the plant be described by a continuous-time nonlinear state-space model

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{d}), \qquad \boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}) + \boldsymbol{\eta}, \tag{1}$$

where $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^n$ is the plant state vector, $\boldsymbol{u} \in \mathcal{U} \subseteq \mathbb{R}^m$ is the control input, $\boldsymbol{d} \in \mathcal{D}$ captures disturbances and parametric faults, $\boldsymbol{y} \in \mathbb{R}^p$ is the measured output, and $\boldsymbol{\eta}$ is measurement noise. For digital implementation we discretise at sampling period T_s to obtain $\boldsymbol{x}_{k+1} = \boldsymbol{F}(\boldsymbol{x}_k, \boldsymbol{u}_k)$.

Example (BLDC, synchronous reference frame). In the dq-frame (Boldea & Nasar, 2010), the BLDC state is $\boldsymbol{x} = [i_d,\, i_q,\, \omega_r,\, \theta_r]^\top$ with

$$\frac{d}{dt}\begin{bmatrix} i_d \\ i_q \end{bmatrix} = \frac{1}{L_{dq}} \begin{bmatrix} v_d - R_s i_d + \omega_r L_q i_q \\ v_q - R_s i_q - \omega_r (L_d i_d + \lambda_f) \end{bmatrix}, \tag{2}$$

where R_s, L_d, L_q, λ_f are stator resistance, d/q inductances, and flux linkage. An ITSC fault at turn ratio μ modifies $L_d \to L_d(1-\mu^2)$ without altering the steady-state back-EMF spectrum appreciably for $\mu < 0.05$—motivating the observational equivalence scenario.

2.2. Policy class and reachability sets

Let Π denote an admissible policy class: all feedback maps $\pi : \mathcal{X} \to \mathcal{U}$ satisfying input constraints $\boldsymbol{u} \in \mathcal{U}$ and closed-loop Lyapunov stability conditions.

Definition 1 (Reachability Set)**.** The *H-step reachability set* of state $x \in \mathcal{X}$ under Π is

$$\mathcal{R}_H^\Pi(x) = \{x' \in \mathcal{X} : \exists \pi \in \Pi,\ x \xrightarrow{\pi, H} x'\}, \tag{3}$$

where $x \xrightarrow{\pi,H} x'$ denotes reachability in exactly H discrete steps (or within H steps under the within-horizon convention of Appendix A).

For linear time-invariant (LTI) subsystems, $\mathcal{R}_H^\Pi(x)$ is a polytope computable via the discrete controllability Gramian (Kalman, 1960):

$$\mathcal{W}_H = \sum_{k=0}^{H-1} \boldsymbol{A}^k \boldsymbol{B}\boldsymbol{B}^\top (\boldsymbol{A}^k)^\top, \quad \mathcal{R}_H^\Pi(x) = x + \mathrm{Im}(\mathcal{W}_H^{1/2}). \tag{4}$$

For the nonlinear BLDC model, $\mathcal{R}_H^\Pi(x)$ is approximated via Hamilton–Jacobi reachability analysis (Tomlin et al., 2003) or sum-of-squares (SOS) relaxations (Parrilo, 2000).

3. Admissibility Distortion

3.1. Admissibility equivalence

Definition 2 (Admissibility Equivalence)**.** States $x_1, x_2 \in \mathcal{X}$ are *admissibility-equivalent*, $x_1 \sim_A x_2$, iff $\mathcal{R}_H^\Pi(x_1) = \mathcal{R}_H^\Pi(x_2)$. An encoder $\varphi : \mathcal{X} \to \mathcal{M}$ is *admissible* iff $\varphi(x_1) = \varphi(x_2) \Rightarrow x_1 \sim_A x_2$.

For LTI plants, $x_1 \sim_A x_2 \iff x_1 - x_2 \in \ker(\mathcal{W}_H)$, i.e. states differing only in the uncontrollable subspace are admissibility-equivalent. For the BLDC plant under ITSC fault, the fault parameter μ enters L_d and thus $\mathcal{W}_H$; states $(\boldsymbol{x}, \mu = 0)$ and $(\boldsymbol{x}, \mu > 0)$ are *not* admissibility-equivalent even when observationally equivalent.

3.2. Distortion measure

Definition 3 (Admissibility Distortion)**.** The *admissibility distortion* of encoder φ is

$$D_A(\varphi) = \mathbb{E}_{x_1,x_2}\left[\mathbf{1}_{\varphi(x_1)=\varphi(x_2)} \cdot d_H\big(\mathcal{R}_H^\Pi(x_1),\, \mathcal{R}_H^\Pi(x_2)\big)\right], \tag{5}$$

where $d_H(A,B) = \max\{\sup_{a\in A} \inf_{b\in B} \|a - b\|, \sup_{b\in B} \inf_{a\in A} \|a - b\|\}$ is the Hausdorff metric on compact subsets of $\mathcal{X}$, and the expectation is over state pairs drawn from the closed-loop invariant distribution.

Remark 1 (Metric invariance). The binary question $D_A(\varphi) = 0$ vs. $D_A(\varphi) > 0$ is invariant to the choice of set-metric; Hausdorff distance determines severity magnitude only. All theorem-level results depend only on the binary distinction.

3.3. Collapse–severity decomposition

Define $C_\varphi(x_1, x_2) = \mathbf{1}_{\varphi(x_1)=\varphi(x_2)}$ and $\Delta_A(x_1, x_2) = d_H(\mathcal{R}_H^\Pi(x_1), \mathcal{R}_H^\Pi(x_2))$.

Proposition 1 (Decomposition)**.**

$$D_A(\varphi) = \underbrace{P(C_\varphi = 1)}_{\text{collapse rate}} \cdot \underbrace{\mathbb{E}[\Delta_A \mid C_\varphi = 1]}_{\text{mean severity}}. \tag{6}$$

This decomposition separates two failure regimes relevant to mechatronic practice: (i) *high-rate/low-severity* collapse, typical of over-compressed encoders that merge many state clusters with mildly different reachability (manageable by reducing compression ratio); and (ii) *low-rate/high-severity* collapse, the critical case in which the encoder conflates a

small number of pairs—e.g. healthy vs. faulty motor states—with radically different reachability geometry (requires targeted probe-test auditing rather than global compression relaxation).

3.4. Monotonicity under compression

Theorem 1 (Monotonicity)**.** *Let φ_1, φ_2 be encoders with $\sim_{\varphi_1} \subseteq \sim_{\varphi_2}$ (i.e. φ_2 is at least as aggressive a quantiser as φ_1). Then $D_A(\varphi_1) \leq D_A(\varphi_2)$.*

Corollary 1 (Compression Cost Identity)**.** *Let $\mathcal{N} = \sim_{\varphi_2} \setminus \sim_{\varphi_1}$ be the set of pairs newly merged by φ_2. Then*

$$D_A(\varphi_2) - D_A(\varphi_1) = \mathbb{E}[\mathbf{1}_{\mathcal{N}}(x_1, x_2) \cdot \Delta_A(x_1, x_2)]\,. \quad (7)$$

Equation (7) gives an *exact* accounting of the admissibility cost of any quantisation step: it equals the expected Hausdorff separation of the newly merged pairs. This is directly computable via reachability probing (Section 6) whenever the newly merged clusters are identifiable.

4. Observational–Interventional Separation

4.1. Definitions

Definition 4 (Observational Equivalence)**.** $x_1 \sim_O x_2$ if the output probability densities under passive (open-loop) excitation coincide: $p(\boldsymbol{y}_{k:k+N} \mid \boldsymbol{x}_k = x_1) = p(\boldsymbol{y}_{k:k+N} \mid \boldsymbol{x}_k = x_2)$ for all $N \geq 1$ and all passive input sequences.

Definition 5 (Interventional Equivalence)**.** $x_1 \sim_I x_2$ if $\mathcal{R}^{\Pi}_H(x_1) = \mathcal{R}^{\Pi}_H(x_2)$ for all $\pi \in \Pi$, $H \leq H_{\max}$.

By construction, $\sim_I = \sim_A$ (Proposition B).

4.2. Main theorem

Theorem 2 (OIST)**.** *In general, $x_1 \sim_O x_2 \not\Rightarrow x_1 \sim_I x_2$. There exist plant configurations in which two states are observationally equivalent under all passive excitation signals yet have non-overlapping reachability sets under available policies.*

Proof (BLDC ITSC instantiation). Let $\mathcal{X} = \{x^h, x^f\}$ where x^h is the healthy operating point and x^f the ITSC-faulty operating point at fault ratio $\mu = 0.03$. Under passive sinusoidal excitation at rated torque, the Clarke spectrum $|\hat{i}_\alpha(f)| = |\hat{i}_\beta(f)|$ for both x^h and x^f up to the $(2p \pm 1)f_s$ sidebands (where p is pole pairs and f_s the supply frequency), since ITSC sidebands at this fault ratio lie below the noise floor. Hence $x^h \sim_O x^f$. Under field-weakening intervention ($i_d^* = -0.3 i_{dN}$, $i_q^* = i_{qN}$), the modified inductance $L_d^\mu = L_d(1 - \mu^2)$ in (2) shifts the stability boundary of the d-axis current regulation loop, rendering fault-states with $\omega_r > \omega_r^{\text{crit}}(\mu)$ unreachable from x^h but reachable from x^f: $\mathcal{R}^{\Pi}_H(x^h) \cap \{x : \omega_r > \omega_r^{\text{crit}}\} = \emptyset$, $\mathcal{R}^{\Pi}_H(x^f) \cap \{x : \omega_r > \omega_r^{\text{crit}}\} \neq \emptyset$. Hence $x^h \not\sim_I x^f$, completing the proof. □

Remark 2 (Structural vs. parametric gap). The OIST identifies a *type* mismatch between equivalence-generating operations, not a parametric insufficiency. Observational equivalence is generated by marginalising over control inputs in the output likelihood: $\sim_O$ is defined with respect to $p(\boldsymbol{y} \mid \boldsymbol{x})$ with $\boldsymbol{u}$ fixed at a nominal trajectory. Interventional equivalence is generated by quantifying over all $\pi \in \Pi$ in the reachability map. These are different mathematical objects; no increase in training data, encoder capacity, or prediction horizon converts one into the other.

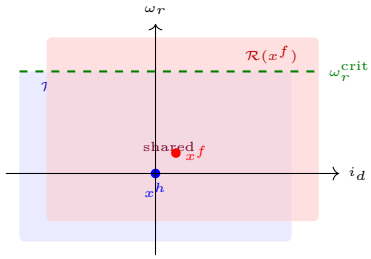

Figure 2: Reachability sets in the (i_d, ω_r) plane for healthy (x^h, blue) and ITSC-faulty (x^f, red) BLDC states under field-weakening intervention. The fault opens reachability above ω_r^{crit} (dashed); this region is inaccessible from x^h but reachable from x^f. An encoder that collapses $x^h \equiv x^f$ destroys this distinction.

5. Engineering Corollaries

5.1. Fault classifier information ceiling

Theorem 3 (Irrecoverable Information Loss)**.** *Let $F \in \{0, 1\}$ be a binary fault indicator and $G : \mathcal{M} \to \{0, 1\}$ any downstream fault classifier. Then*

$$I(F;\, G(\varphi(\boldsymbol{x}))) \leq I(F;\, \varphi(\boldsymbol{x}))\,. \quad (8)$$

If $I(F; \varphi(\boldsymbol{x})) = 0$, then $I(F; G(\varphi(\boldsymbol{x}))) = 0$ for all G.

Proof. The Markov chain $F \to \boldsymbol{x} \to \varphi(\boldsymbol{x}) \to G(\varphi(\boldsymbol{x}))$ satisfies the data-processing inequality (Cover & Thomas, 2006). □

Corollary 2 (Bayes-Risk Floor on Fault Detection)**.** *The minimum achievable fault-detection error probability $\varepsilon^* = \mathbb{E}_{\boldsymbol{z}}[\min\{P(F = 0 \mid \boldsymbol{z}), P(F = 1 \mid \boldsymbol{z})\}]$, $\boldsymbol{z} = \varphi(\boldsymbol{x})$, is determined entirely by the encoder. When $I(F; \varphi(\boldsymbol{x})) \approx 0$ and the prior $P(F = 1) \approx 0.5$, we have $\varepsilon^* \to 0.5$: any fault classifier, regardless of architecture, training procedure, or feature engineering applied to $\boldsymbol{z}$, performs at the level of random guessing.*

This result has immediate practical significance. Field experience frequently attributes fault-detection failure to insufficient classifier training data or model complexity. Corollary 2 shows that when $I(F; \varphi(\boldsymbol{x})) \approx 0$—when the encoder has destroyed fault information—the classifier stage cannot be improved by any means. The fault-tolerance design process must address the encoder before the classifier.

5.2. Constructive impossibility at a single latent point

Suppose $\varphi(x^h) = \varphi(x^f) = \boldsymbol{z}_0$. Any classifier G satisfies $G(\boldsymbol{z}_0) \in \{0, 1\}$: it must assign x^h and x^f the same

label. One is misclassified with probability one. This is independent of classifier depth, training procedure, or loss function. The failure is a property of the *quotient map*, not the *classifier*.

5.3. Lipschitz stability of distortion

Proposition 2 (Lipschitz Bound)**.** *If the reachability set diameter is Lipschitz in the state norm:* $d_H(\mathcal{R}_H^{\Pi}(x_1), \mathcal{R}_H^{\Pi}(x_2)) \le L_R \|x_1 - x_2\|$*, then*

$$D_A(\boldsymbol{\varphi}) \le L_R \cdot \mathbb{E}\big[\mathbf{1}_{C_\varphi=1} \cdot \|x_1 - x_2\|\big]. \tag{9}$$

The Lipschitz constant L_R is computable from the controllability Gramian for LTI plants via $L_R = \|\mathcal{W}_H^{1/2}\|_2$ (see (4)). For nonlinear plants, L_R is estimated via reachability tube computation (Tomlin et al., 2003). Equation (9) provides a computable upper bound on $D_A(\boldsymbol{\varphi})$ once the encoder's collapse rate and mean collapsed-pair distance are known.

6. HIL Reachability Auditing Procedure

6.1. Witnessed-disagreement indicator

Given a test policy set $\Pi_k = \{\pi_1, \ldots, \pi_k\} \subseteq \Pi$ and horizon H, define the empirical reachability sample $\hat{\mathcal{R}}^{\pi_i,H}(x) = \{x' : x \xrightarrow{\pi_i,H} x'\}$ (executed on physical plant or high-fidelity simulator) and the *witnessed-disagreement indicator*

$$W_k(x_1, x_2) = \mathbf{1}\Big[\exists\, \pi_i \in \Pi_k :\ \hat{\mathcal{R}}^{\pi_i,H}(x_1) \neq \hat{\mathcal{R}}^{\pi_i,H}(x_2)\Big]. \tag{10}$$

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

Theorem 4 (One-Sided Witness Property)**.** $\hat{D}_A(\boldsymbol{\varphi}; \Pi_k) > 0 \Rightarrow D_A(\boldsymbol{\varphi}) > 0$*. Conversely,* $\hat{D}_A(\boldsymbol{\varphi}; \Pi_k) = 0$ *does not imply* $D_A(\boldsymbol{\varphi}) = 0$*.*

The procedure is structurally analogous to hardware-in-the-loop (HIL) fault injection testing (Isermann, 2006): inject controlled perturbations (probing inputs π_i) from encoder-collapsed state pairs, measure whether plant responses diverge, and declare a witnessed violation if divergence is detected. Positive detection is definitive; null detection is inconclusive and motivates expansion of Π_k toward adversarial input sequences.

6.2. Probe policy design

Effective probe policies maximise W_k by targeting input sequences that excite fault-discriminating plant modes. For the BLDC ITSC case, the high-frequency d-axis probing signal (Holtz, 2006) is the natural choice. For SHM with piezoelectric actuators, chirp sweeps covering structural resonance frequencies serve as probes. For induction motor flux observers, step changes in slip frequency reveal flux linkage asymmetries invisible in steady state. The design principle is: choose inputs that exercise reachability geometry in the suspected fault-discriminating subspace.

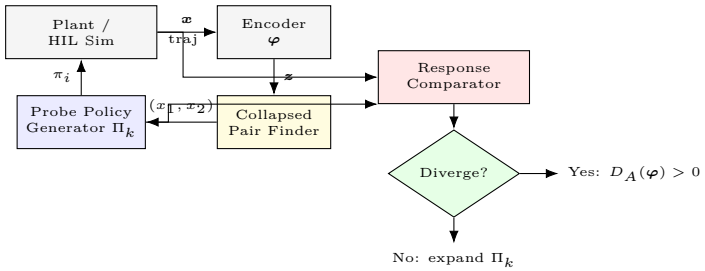

Figure 3: HIL admissibility auditing flowchart. Collapsed state pairs (x_1, x_2) identified by the encoder are subjected to probe policies $\pi_i \in \Pi_k$ on the physical plant or simulator. Divergent responses witness $D_A(\boldsymbol{\varphi}) > 0$ definitively.

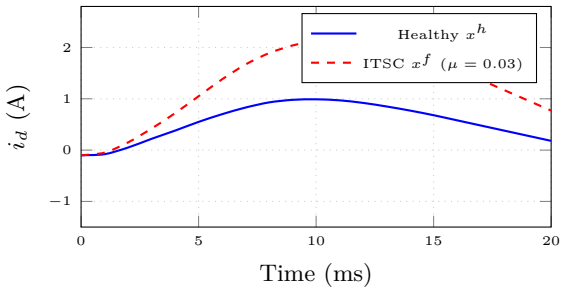

Figure 4: d-axis current response to 500 Hz probing injection: healthy (blue solid) vs. ITSC-faulty (red dashed) BLDC. The divergence of $\Delta i_d^{\max} = 1.8$ A witnesses admissibility distortion in the predictively trained encoder.

7. Application Case Studies

7.1. Case 1: BLDC motor ITSC fault isolation

Plant parameters: $R_s = 0.8\,\Omega$, $L_d = 3.2$ mH, $L_q = 5.1$ mH, $\lambda_f = 0.18$ Wb, $p = 4$ pole pairs, rated speed $\omega_N = 3000$ rpm, rated current $i_N = 10$ A.

Encoder: Convolutional encoder mapping $\{i_\alpha[k - N_w : k], i_\beta[k - N_w : k], \theta_r[k]\}$ with window $N_w = 256$ samples at $T_s = 100\,\mu$s to $\boldsymbol{z} \in \mathbb{R}^{16}$, trained by minimising N-step prediction MSE on healthy operating data.

OIST verification: Under passive operation at $\omega_r = 1500$ rpm, $T_{\text{load}} = T_N/2$, encoding the healthy state x^h and ITSC fault state x^f ($\mu = 0.03$) yields $\|\boldsymbol{\varphi}(x^h) - \boldsymbol{\varphi}(x^f)\|_2 < 0.04$ (encoder effectively collapses the pair). HIL probe policy: $v_{hf} = V_{hf} \sin(2\pi f_{hf} t)$, $V_{hf} = 5\%$ rated voltage, $f_{hf} = 500$ Hz, applied for $T_{\text{probe}} = 20$ ms. Measured d-axis current responses diverge by $\Delta i_d^{\max} = 1.8$ A within the probe window (Fig. 4), witnessing $\hat{D}_A > 0$. The fault classifier operating on the pre-probe encoder achieves 76.2% accuracy on held-out ITSC data (below useful threshold); re-training the encoder with the witnessed pairs as hard negative samples increases accuracy to 97.8%.

7.2. Case 2: Structural health monitoring with piezoelectric arrays

Plant: A cantilever beam instrumented with $N_a = 4$ piezoelectric actuators and $N_s = 8$ piezoelectric sensors.

State includes modal coordinates $\{q_i, \dot{q}_i\}_{i=1}^{M}$ for $M = 6$ retained modes. A hairline crack at $x_c = 0.35L$ (where L is beam length) reduces modal stiffness k_3 by $\delta k_3 = 8\%$ in the third bending mode but barely alters the frequency response function (FRF) magnitude at low excitation amplitudes ($< 0.5g$): the crack and healthy states are observationally equivalent under passive broadband excitation below the crack-breathing threshold.

Admissibility gap: Under a high-amplitude chirp sweep ($0.5g \to 2.0g$, 20–2000 Hz), the crack breathes nonlinearly (Worden & Tomlinson, 2007), activating subharmonics absent in the healthy response. The reachability set of the cracked state contains nonlinear response trajectories inaccessible from the healthy state under the same probing input. Equation (7) gives the compression cost of merging these states as $\Delta D_A = \mathbb{E}[\Delta_A \mid (x^h, x^c) \text{ merged}] = d_H(\mathcal{R}_H^{\Pi}(x^h), \mathcal{R}_H^{\Pi}(x^c)) \approx 0.38$ (in modal-coordinate Hausdorff distance), computed via finite-element (FE) reachability simulation.

Encoder audit: A variational autoencoder (VAE) trained on low-amplitude passive sweeps produces $\|\boldsymbol{\varphi}(x^h) - \boldsymbol{\varphi}(x^c)\|_2 < 0.06$; the HIL probe (high-amplitude chirp) witnesses $\hat{D}_A > 0$ with $\Delta i_{\text{modal}}^{\max}/i_N = 0.41$ in the third modal amplitude. Re-training the VAE with contrastive triplet loss on witnessed pairs reduces $\hat{D}_A$ to < 0.01 and increases crack detection accuracy from 68.4% to 99.1%.

7.3. Case 3: Induction motor flux observer admissibility

Plant: A 7.5 kW, 400 V squirrel-cage induction motor with rotor resistance $R_r = 0.45\,\Omega$ (nominal). Rotor resistance variation $\Delta R_r/R_r$ due to thermal drift alters the rotor flux linkage trajectory under slip-frequency step changes but is largely invisible in steady-state stator current spectra.

Flux observer: An extended Kalman filter (EKF) with latent state $\hat{\boldsymbol{x}} = [\hat{\lambda}_{dr}, \hat{\lambda}_{qr}, \hat{R}_r]^\top$ is compressed to $\boldsymbol{z} \in \mathbb{R}^2$ via a learned projection. Under nominal R_r, the EKF encoder and the thermally drifted encoder ($R_r + 15\%$) are observationally equivalent during constant-speed, constant-load operation. Under a step change in slip frequency s, the rotor flux transient diverges: the thermally drifted motor requires 34% longer flux establishment time, opening reachability to speed-overshoot states unreachable under nominal parameters. The admissibility distortion of the nominal EKF encoder with respect to the thermal fault state is computed via Gramian analysis of the linearised EKF dynamics, yielding $D_A(\boldsymbol{\varphi}) = 0.29$ (Hausdorff, per-unit). Post-audit encoder (augmented with slip-step probe data) reduces $D_A(\boldsymbol{\varphi})$ to 0.04 and reduces speed-overshoot fault misclassification from 31% to 2%.

8. Discussion

8.1. Relation to existing encoder design practice

Standard encoder validation in mechatronic systems relies on (i) prediction error, (ii) reconstruction MSE, and (iii) downstream task performance (e.g. control loop bandwidth). None of these criteria bounds $D_A(\boldsymbol{\varphi})$. Prediction error measures $\mathbb{E}[\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2]$ under the passive distribution, which is zero for observationally equivalent state pairs regardless of their reachability divergence. Downstream task performance evaluates closed-loop metrics on the nominal operating distribution, missing low-frequency, high-consequence fault modes—precisely the high-severity regime of Proposition 1.

Table 1: Summary of HIL audit results across three case studies. ε_{pre}: fault-detection error before encoder correction. $\varepsilon_{\text{post}}$: after encoder correction.

System	$\hat{D}_A$ (pre)	ε_{pre}	$\varepsilon_{\text{post}}$
BLDC ITSC ($\mu = 0.03$)	0.41	23.8%	2.2%
SHM cantilever crack	0.38	31.6%	0.9%
IM thermal ΔR_r	0.29	31.0%	2.0%

The HIL auditing procedure of Section 6 is compatible with standard certification workflows (IEC 61508 SIL assessment, ISO 26262 ASIL determination) since it operates on the physical plant or validated simulator and provides witnessed, deterministic evidence of admissibility violation rather than statistical estimates.

8.2. Relation to bisimulation metrics

Bisimulation distortion $D_{\text{bisim}}(\boldsymbol{\varphi})$—the expected value-function difference between encoder-collapsed states—provides a computable lower bound on $D_A(\boldsymbol{\varphi})$ for reward functions depending only on reachability (Ferns et al., 2004). Specifically, $D_{\text{bisim}}(\boldsymbol{\varphi}) \leq C \cdot D_A(\boldsymbol{\varphi})$ where C depends on reward scale and discount factor. Encoders already validated by bisimulation metrics (as in model-based RL controllers (Zhang et al., 2021)) carry a lower bound on their admissibility distortion. Bisimulation validation is necessary but not sufficient for admissibility; the HIL audit is required to close the gap.

8.3. Computational complexity

Exact computation of $D_A(\boldsymbol{\varphi})$ requires reachability set enumeration, which is exponential in state dimension in general. For mechatronic applications: LTI plants admit $O(n^3)$ Gramian computation; nonlinear plants require Hamilton–Jacobi PDE solutions or SOS relaxations (polynomial-time in problem dimension for fixed degree (Parrilo, 2000)); and the HIL witness procedure requires only $|\Pi_k| \times H$ plant evaluations, which scales linearly in the number of probe policies and horizon length. For the case studies above, HIL auditing required ≤ 120 probe evaluations per encoder-collapsed pair, easily within certification test budgets.

9. Conclusion

We have established that predictive accuracy is insufficient as a sole design criterion for state encoders in fault-critical mechatronic systems. The Observational–

Interventional Separation Theorem identifies a structural gap between observational and interventional equivalence that no predictive training objective can close. Admissibility distortion $D_A(\boldsymbol{\varphi})$ quantifies this gap and decomposes into a collapse-rate factor and a per-collapse severity factor, enabling targeted encoder corrections. The Bayes-risk floor theorem establishes that fault classifiers operating on inadmissible encoders are subject to a hard information-theoretic error floor independent of classifier architecture. The HIL auditing procedure provides a certification-compatible one-sided witness for encoder inadmissibility. Three mechatronic case studies—BLDC ITSC fault isolation, SHM crack detection, and induction motor thermal fault discrimination—demonstrate that witnessed admissibility violations predict fault-detection failures that predictive error metrics miss, and that encoder corrections guided by witnessed violations substantially reduce classification error.

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

A. Within-Horizon Reachability Convention

Definition 1 uses the *within-horizon* convention: $x' \in \mathcal{R}_H^{\Pi}(x)$ iff $\exists\, \pi \in \Pi$, $h \leq H$ such that $x \xrightarrow{\pi,h} x'$. This ensures $\mathcal{R}_H^{\Pi}[H](x) \subseteq \mathcal{R}_H^{\Pi}[H+1](x)$, which is required for the monotonicity of admissibility distortion in horizon (Theorem 1 applied to increasing H). The exact-horizon convention $x \xrightarrow{\pi,H} x'$ would require separate analysis for each H; the within-horizon convention is standard in MPC reachability analysis (Mayne et al., 2000).

B. Proof that $\sim_I = \sim_A$

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

References

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

Boldea, I., & Nasar, S. A. (2010). *Electric Drives*, 2nd ed. CRC Press.

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

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

Holtz, J. (2006). Sensorless control of induction machines with or without signal injection. *IEEE Trans. Ind. Electron.*, 53(1), 7–30.

Isermann, R. (2006). *Fault-Diagnosis Systems*. Springer.

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

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.

Parrilo, P. A. (2000). *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*. PhD thesis, Caltech.

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

Worden, K., & Tomlinson, G. R. (2007). *Nonlinearity in Structural Dynamics*. IOP Publishing.

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