*Proceedings of the 11th North American International Conference on Industrial Engineering and Operations Management, Milwaukee, Wisconsin, USA, June 9-11, 2026*

*Proceedings of the International Conference on Industrial Engineering and Operations Management*

Publisher: IEOM Society International, USA DOI: 10.46254/NA11.20260043
Published: June 9, 2026

# Preserving Behavioral Coherence in Reliability-Aware Neural Decision Systems

**Mayra Bornacelly**
PhD Student
Department of Industrial Engineering and Management Systems
University of Central Florida
Orlando, FL, USA
mayra.bornacellycastaneda@uf.edu

## Abstract

Auxiliary control variables such as gating, routing, and reliance coefficients are widely used to arbitrate competing information sources in neural systems, yet they are typically optimized only indirectly through downstream task losses. As a result, these variables can collapse into low-sensitivity, low-variance, or behaviorally inconsistent regimes while nominal predictive performance remains high. We introduce function-space behavioral priors that preserve the functional role of a reliance variable $\alpha$ by constraining the local sensitivity of the decision function with respect to $\alpha$. An effect prior prevents functional suppression through a reliability-weighted sensitivity floor, while a monotonicity prior promotes directional consistency with internal arbitration signals. We formulate training as a hybrid MAP objective combining task likelihood, Bayesian regularization, and derivative-based behavioral constraints, and provide local guarantees showing that the priors penalize collapse-prone regimes. Across two domains, perturbation settings, baseline comparisons, and ablations, the priors improve alignment between $\alpha$ and retrieval reliability, stabilize control dynamics, and improve robustness and predictive performance under higher distributional complexity. These results suggest that preserving behavioral coherence in internal control variables can serve as a principled inductive bias for reliability-aware neural decision systems.

## Keywords
Reliability-Aware Neural Decision Systems, Function-Space Behavioral Regularization, Robust Machine Learning, Memory-Augmented Decision Support and Behavioral Coherence in AI Systems.

## 1. Introduction
Neural decision systems increasingly rely on auxiliary control variables to arbitrate between competing information sources, including retrieval, memory, uncertainty, and expert pathways. Yet these variables are typically supervised only indirectly through task loss, making them susceptible to behavioral collapse: they may become low-sensitivity, low-variance, or directionally inconsistent while predictive performance remains high. Existing regularizers (weight decay, dropout, Bayesian priors, or architectural constraints) primarily act in parameter space and therefore do not directly preserve the functional role of such variables. Related failures such as posterior collapse and attention entropy collapse suggest that optimization can silently degrade internal mechanisms, motivating regularization in function space rather than weight space. We study this problem through a reliance variable $\alpha$ that arbitrates between current perceptual evidence and retrieved memory in a memory-augmented Bayesian decision system. We introduce lightweight function-space behavioral priors that constrain the local sensitivity of the decision function with respect to $\alpha$. An effect prior enforces reliability-weighted functional relevance, while a monotonicity prior enforces directional consistency with internal arbitration signals. Together, these priors bias optimization away from collapse-prone regimes without constraining $\alpha$ directly or modifying the architecture. We formalize collapse regimes for auxiliary control variables and provide local guarantees showing that the proposed priors reduce low-sensitivity and

directionally inconsistent regimes by reshaping the local optimization geometry. Across two domains, perturbation settings, baseline comparisons, and ablations, the priors improve alignment between $\alpha$ and retrieval reliability, stabilize control dynamics, and yield stronger robustness and predictive performance under higher distributional complexity.

### 1.1 Objectives

The objective of this work is to develop and evaluate function-space behavioral priors that preserve the functional role of auxiliary reliance variables in neural decision systems. Specifically, we aim to prevent low-sensitivity and directionally inconsistent collapse of the reliance coefficient α, improve its alignment with retrieval reliability, and assess whether preserving internal behavioral coherence enhances robustness, uncertainty reduction, and predictive performance under distributional complexity.

## 2. Literature Review

Collapse phenomena in neural systems show that internal mechanisms can become uninformative under optimization even when task performance remains strong. In variational latent-variable models, posterior collapse occurs when the approximate posterior approaches the prior, causing the decoder to ignore the latent representation (Lucas et al. 2019). Analogous failures occur in attention mechanisms, where entropy collapse concentrates attention mass on a small number of tokens, reducing effective information integration and destabilizing training (Zhai et al. 2023). These cases suggest that optimization can silently degrade internal mechanisms. However, existing collapse analyses primarily focus on latent representations or attention distributions rather than auxiliary control variables that arbitrate between competing evidence sources. Classical regularization methods primarily operate in parameter space. Weight decay penalizes parameter magnitude to improve generalization (Krogh and Hertz 1991), dropout reduces co-adaptation by randomly omitting units during training (Srivastava et al. 2014), and Bayesian neural networks place priors over weights to control model complexity and epistemic uncertainty (MacKay 1992; Neal 1996). These methods constrain capacity or uncertainty, but do not directly constrain the functional role of internal control variables. This limitation is central in reliance-based systems, where gating variables, memory arbitration coefficients, or expert-routing weights must remain behaviorally meaningful rather than merely numerically optimized. Furthermore, function-space regularization offers a more behavior-centered alternative by placing priors over predictive functions rather than parameters (Rudner et al. 2023). This perspective is closely aligned with the goal of preserving what a model does, rather than merely constraining how its parameters are configured. However, existing function-space regularization frameworks typically encode global functional preferences over the predictive function. In contrast, this work defines targeted derivative-based priors that regulate the local sensitivity of the decision function with respect to an auxiliary reliance variable $\alpha$, yielding a mechanism-specific function-space regularization strategy for preserving evidence-gating behavior under optimization. Moreover, related work on monotonic and structured neural constraints further shows that directional relationships can be encoded as useful inductive biases. Monotone and partially monotone neural networks, for example, preserve directional effects while retaining representational flexibility (Daniels and Velikova 2010). However, such approaches generally impose global input-output constraints or architectural restrictions. The monotonicity prior proposed here instead operates locally in function space, constraining only the directional influence of $\alpha$ relative to an internal arbitration signal derived from memory-informed and attention-only logits. Finally, reliability and calibration research has shown that neural confidence estimates may be poorly aligned with predictive correctness, motivating post-hoc methods such as temperature scaling (Guo et al. 2017). In this work, reliability is used differently: not as an output-level calibration target, but as an internal gating signal that modulates the strength of the effect prior. Thus, reliability estimation is connected to preserving internal mechanism behavior rather than merely improving confidence interpretation. Existing work analyzes collapse, parameter regularization, global functional constraints, or output calibration, but does not directly address auxiliary control-variable collapse as a local function-space failure mode; we fill this gap with derivative-based behavioral priors that actively preserve the relevance and directional consistency of $\alpha$ under optimization.

## 3. Methods

### 3.1 Alpha Construction

We consider a reliance variable $\alpha \in [0,1]$ that modulates the contribution of retrieved memory relative to current perceptual evidence. Given an attention-derived representation $z_{attn}$ and a retrieved memory representation $z_{mem}$, the final decision logit is computed as $f(x, \alpha) = (1 - \alpha)z_{attn} + \alpha z_{mem}$. The value of $\alpha$ is inferred as an instance-dependent stochastic gating coefficient using a learned neural head parameterizing a Beta distribution conditioned on latent similarity signals and retrieval-related cues. The sampled value is further modulated by an internal reliability estimate $R(x)$, yielding a reliability-aware arbitration mechanism between current evidence and memory retrieval. In this work, we treat the construction of $\alpha$ as given and focus on how optimization reshapes its functional role under distributional complexity and shift.

### 3.2 Function-Space Behavioral Priors

Auxiliary control variables in neural systems are vulnerable to collapse under unconstrained optimization, often converging toward weakly-informative, low-variance, or behaviorally inconsistent regimes despite strong nominal task performance. Existing regularization methods typically operate in parameter space and therefore do not directly constrain the functional role of such variables. To address this, we introduce lightweight function-space behavioral priors that regularize how the decision function responds to the reliance variable $\alpha$. Rather than constraining parameter values, these priors impose structural preferences on the local sensitivity of the predictive function $f(x, \alpha)$, encouraging $\alpha$ to remain both functionally relevant and directionally consistent. Concretely, we define two complementary constraints: (i) an effect prior, which prevents low-sensitivity collapse by enforcing a minimum functional dependence on $\alpha$, and (ii) a monotonicity prior, which encourages directional alignment between the influence of $\alpha$and internal arbitration signals. Together, these priors act as behavioral regularizers, shaping the local geometry of optimization without imposing rigid architectural restrictions.

**3.2.1 MAP Training Objective.** We formulate training as a hybrid Bayesian–deterministic Maximum A Posteriori (MAP) objective that combines task likelihood, epistemic regularization, and function-space behavioral priors: $L = L_{NLL} + \beta\, KL(q(\theta) \parallel p(\theta)) + \lambda_1 L_{mono} + \lambda_2 L_{effect}$. here, $L_{NLL}$ denotes the negative log-likelihood of the observed labels, and the KL term regularizes the variational posterior $q(\theta)$ of the Bayesian decision layer toward its prior $p(\theta)$, corresponding to a partial ELBO objective. The upstream encoder, memory mechanism, and reliance parameter $\alpha$ are optimized as point estimates. The coefficients $\beta$, $\lambda_1$, and $\lambda_2$ control the relative strength of epistemic and behavioral regularization. From a probabilistic perspective, $\lambda_1$and $\lambda_2$ admit a MAP interpretation as prior strengths over desired functional behaviors of the decision function.

**3.2.2 Formalization of Collapse Regimes.** We formalize collapse as the degeneration of the functional role of the reliance variable $\alpha$ into undesirable operating regimes. Specifically, we define three forms of collapse. Low-sensitivity collapse occurs when the decision function becomes locally insensitive to $\alpha$:

$$C_{low} = \left\{x: \left| \frac{\partial f(x,\alpha)}{\partial \alpha} \right| < \epsilon \right\},$$

where $\epsilon > 0$ is a small threshold. In this regime, $\alpha$ has negligible functional effect. Directional collapse occurs when the influence of $\alpha$ becomes inconsistent with internal arbitration signals:

$$C_{dir} = \left\{x: \frac{\partial f(x,\alpha)}{\partial \alpha}\, u(x) < 0 \right\},$$

where $u(x)$ denotes the target direction of influence. In this regime, increasing reliance produces behavior opposite to the model's internal ordering preference. Variance collapse occurs when $\alpha$ contracts toward a near-constant operating point: $C_{var} = \{Var(\alpha) < \delta\}$, for a small $\delta > 0$. This regime reflects loss of adaptive modulation across inputs. The proposed function-space priors are designed to reduce the measure of these collapse sets under the training distribution by preserving functional relevance, directional consistency, and structured variability.

**3.2.3 Monotonicity Prior.** To preserve directional consistency in the influence of $\alpha$, we introduce a monotonicity prior that encourages the decision function to move in agreement with internal arbitration signals:

$$L_{mono} = E_x \left[ softplus \left( -\frac{\partial f(x,\alpha)}{\partial \alpha}\, u(x) \right) \right],$$

Where $u(x) = sign(z_{mem} - z_{attn})$ defines the target direction of influence using only sign information. For the decision function $f(x,\alpha) = (1-\alpha) z_{attn} + \alpha z_{mem}$, the sensitivity with respect to $\alpha$ is

$$\frac{\partial f(x,\alpha)}{\partial \alpha} = z_{mem} - z_{attn}.$$

Thus, the prior penalizes instances in which increasing reliance shifts the decision in a direction inconsistent with the model's internal ordering preference.

**Proposition 1 (Directional Consistency).** Assume $f(x,\alpha)$ is continuously differentiable in $\alpha$, and $u(x)$ provides a locally meaningful directional signal. Then minimizing $L_{mono}$ penalizes the directional collapse set and encourages a decrease of the directional collapse set $C_{dir} = \left\{x: \frac{\partial f(x,\alpha)}{\partial \alpha} u(x) < 0\right\}$ under the training distribution.

**Justification.** The softplus penalty is monotonically decreasing in $\frac{\partial f}{\partial \alpha} u(x)$, and approaches zero as the product becomes positive and large. Gradient descent therefore increases this product in expectation, reducing the probability of sign inconsistency.

**3.2.4 Effect Prior.** While the monotonicity prior constrains the direction of influence, it does not prevent the model from suppressing the effect of $\alpha$altogether. To preserve functional relevance, we introduce an effect prior that enforces a minimum sensitivity of the decision function to $\alpha$ when memory retrieval is reliable: $L_{effect} = E_x \left[softplus\left(m(x) - \| \frac{\partial f(x,\alpha)}{\partial \alpha} \|^2\right)\right]$, where $m(x) = m_0 R(x)$, $m_0 > 0$ is a global sensitivity floor, and $R(x) \in [0,1]$ is an internal retrieval reliability signal. Thus, the prior becomes active only when retrieved memory is expected to be informative. This prior acts as a soft lower-bound constraint on the squared sensitivity of the decision function, discouraging low-sensitivity collapse without imposing hard feasibility requirements.

**Proposition 2 (Sensitivity Preservation).** Assume $f(x,\alpha)$ is continuously differentiable in $\alpha$, and $R(x) \in [0,1]$. Then minimizing $L_{effect}$ encourages an increase in expected squared sensitivity $E\left[\| \frac{\partial f(x,\alpha)}{\partial \alpha} \|^2\right]$ and decreases the expected measure of the low-sensitivity collapse set $C_{low} = \left\{x: | \frac{\partial f(x,\alpha)}{\partial \alpha} | < \epsilon\right\}$. In expectation, the induced sensitivity is encouraged toward a reliability-weighted lower bound proportional to $m_0 E[R(x)]$**.**

**Justification.** The softplus penalty is monotonically decreasing in the squared sensitivity term. Gradient descent therefore increases $\| \frac{\partial f}{\partial \alpha} \|^2$ when it falls below $m(x)$, while naturally attenuating the constraint when reliability is low.

**3.2.5 Local Optimization Geometry and Stability**

Taken together, the monotonicity and effect priors alter the local optimization geometry of the training objective by penalizing degenerate functional regimes of $\alpha$. In particular, they reshape the gradient field around stationary points where the reliance signal becomes weakly-informative or behaviorally inconsistent. Let the parameter update follow gradient descent: $\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta_t)$, where $L = L_{NLL} + \beta\, KL + \lambda_1 L_{mono} + \lambda_2 L_{effect}$. The behavioral priors contribute additional gradient components that explicitly depend on $\frac{\partial f}{\partial \alpha}$, thereby modifying the local dynamics of optimization in regions where $\alpha$becomes inactive or misaligned.

**Proposition 3 (Destabilization of Degenerate Stationary Points).** Assume $f(x,\alpha)$ is twice continuously differentiable, and optimization proceeds via sufficiently small gradient steps. Then stationary points satisfying either $| \frac{\partial f(x,\alpha)}{\partial \alpha} | \approx 0$ or $\frac{\partial f(x,\alpha)}{\partial \alpha} u(x) < 0$ incur non-zero gradients from $L_{effect}$or $L_{mono}$, respectively. Consequently, such points are no longer stationary under the augmented objective unless compensated by the task likelihood.

**Justification.** At low-sensitivity points, the effect prior contributes gradients that increase $\| \frac{\partial f}{\partial \alpha} \|^2$. At directionally inconsistent points, the monotonicity prior contributes gradients that increase $\frac{\partial f}{\partial \alpha} u(x)$**.** Thus, the priors perturb the local gradient field away from degenerate regimes, making collapse-prone equilibria less stable under optimization.

**3.2.6 Assumptions and Limitations.** The preceding propositions provide local guarantees on the behavior of the proposed priors under standard smoothness assumptions. In particular, these results rely on the following conditions:

**A1 (Differentiability).** The decision function $f(x, \alpha)$ is continuously differentiable in $\alpha$, and twice differentiable when considering local optimization geometry.

**A2(Meaningful Direction Signal).** The internal arbitration signal $u(x) = sign(z_{mem} - z_{attn})$ provides a locally meaningful ordering preference. If this signal is noisy or systematically biased, the monotonicity prior may enforce self-consistent but semantically suboptimal behavior.

**A3 (Calibrated Reliability Signal).** The retrieval reliability estimate $R(x) \in [0,1]$ provides a meaningful proxy for memory informativeness. Miscalibrated reliability estimates may incorrectly activate or suppress the effect prior.

**A4 (Small-Step Optimization).** The local stability interpretation assumes sufficiently small optimization steps such that first-order local geometry remains informative across updates. Under these assumptions, the proposed priors reduce the measure of collapse-prone regimes and reshape local optimization dynamics. However, they do not guarantee global optimality, global avoidance of collapse, or semantic correctness of internal arbitration signals. The guarantees are therefore best interpreted as local inductive biases that improve the stability and functional coherence of the reliance mechanism during training.

### 3.3 Implementation Details

In all experiments, the Bayesian decision layer is trained using a β-weighted KL regularization term corresponding to a partial β-ELBO objective, with $\beta = 0.01$. This provides mild epistemic regularization without dominating the task likelihood or interacting adversely with the function-space priors. Unless otherwise specified, the monotonicity and effect priors are weighted equally, with fixed coefficients $\lambda_1 = \lambda_2 = 1$. These values were held constant across domains and model variants, and no additional hyperparameter tuning was performed. The proposed priors require only first-order derivatives with respect to the scalar control variable $\alpha$, and do not require Hessian computations or Jacobian–vector products. As a result, the computational overhead is negligible relative to standard backpropagation and scales linearly with batch size.

## 4. Data Collection and Dataset Characteristics

This study builds on the same two-domain dataset introduced in our prior work (Bornacelly Castaneda 2025), consisting of AI-generated texts labeled as 0 and human-written texts labeled as 1. Both domains were preprocessed using the same tokenization pipeline, converting texts into vocabulary-index sequences with a maximum vocabulary size of 5,000. Domain 1 provides a balanced classification setting, with 9,750 samples per class and relatively short sequences. Human-written texts average approximately 27 tokens, while AI-generated texts average approximately 52 tokens. In contrast, Domain 2 introduces a more challenging distributional setting, with 12,750 AI-generated samples and 2,150 human-written samples. Texts in this domain are substantially longer, averaging approximately 133–175 tokens, and exhibit greater lexical variability and token repetition. After preprocessing, the longest retained sequence contains 238 tokens in Domain 1 and 384 tokens in Domain 2. Thus, Domain 1 serves as a balanced baseline dataset, whereas Domain 2 provides a more complex benchmark characterized by class imbalance, longer sequences, higher vocabulary diversity, and increased intra-class variability due to AI texts generated by multiple models.

## 5. Results and Discussion

### 5.1 Empirical Validation of Behavioral Regularization

We first verify whether the proposed priors reshape the behavior of the reliance signal α. Across both domains, α becomes more strongly aligned with cosine similarity and retrieval reliability after regularization, with the strongest change in Domain 2, where α–reliability correlation increases from approximately 0.20 to 0.73 (Table 1). The priors also shift α toward more structured operating regimes and preserve variability throughout training These behavioral changes coincide with modest improvements in predictive performance and reduced predictive uncertainty suggesting that the priors act as behavioral regularizers rather than merely interpretability diagnostics. Supplementary distributional statistics and trajectory plots are provided in Appendix A.

### 5.2 Baseline Comparison

To assess whether the observed gains arise from the proposed behavioral priors rather than regularization alone, we compare against representative parameter-space regularization strategies: Bayesian Only, which includes Bayesian priors over network weights without behavioral constraints, and Weight Decay, which adds classical L2 regularization. The Proposed model incorporates both the effect prior and monotonicity prior. All methods are evaluated across both domains and multiple random seeds, and compared using predictive performance (Accuracy, F1, ROC-AUC) and uncertainty metrics. This experiment isolates whether explicitly constraining decision sensitivity yields benefits beyond conventional parameter-space regularization. Table 1 summarizes performance across representative regularization strategies. In Domain 1, all methods perform similarly, indicating a largely saturated task. Bayesian-only and weight-decay achieve marginally higher predictive metrics, while the proposed priors maintain competitive performance with lower variance, suggesting improved optimization stability. In Domain 2, the proposed method consistently outperforms all baselines, achieving the highest Accuracy (0.856 ± 0.002), F1-score (0.854 ± 0.003), and ROC-AUC (0.915 ± 0.002), while also reducing predictive uncertainty. Notably, standard weight decay exhibits severe instability in Domain 2, with high variance across seeds, suggesting that parameter-space regularization alone may over-constrain learning under more complex distributional conditions. These results indicate that behavioral constraints on decision sensitivity provide limited gains in easier settings, but become increasingly beneficial as task complexity and optimization pressure increase.

Table 1. Baseline comparison across domains

| Domain | Method | Accuracy ↑ | F1 ↑ | ROC-AUC ↑ | Uncertainty ↓ |
|---|---|---|---|---|---|
| D1 | Bayesian Only | 0.953 ± 0.005 | 0.954 ± 0.005 | 0.984 ± 0.000 | 0.021 ± 0.009 |
| D1 | Weight Decay | 0.953 ± 0.003 | 0.954 ± 0.003 | 0.983 ± 0.001 | 0.024 ± 0.008 |
| D1 | Proposed | 0.951 ± 0.001 | 0.952 ± 0.001 | 0.982 ± 0.001 | 0.021 ± 0.006 |
| D2 | Bayesian Only | 0.830 ± 0.015 | 0.833 ± 0.019 | 0.901 ± 0.010 | 0.012 ± 0.001 |
| D2 | Weight Decay | 0.580 ± 0.197 | 0.690 ± 0.102 | 0.590 ± 0.251 | 0.012 ± 0.005 |
| D2 | Proposed | 0.856 ± 0.002 | 0.854 ± 0.003 | 0.915 ± 0.002 | 0.009 ± 0.002 |

## 5.3 Robustness Evaluation

To assess whether the proposed function-space priors preserve the behavioral role of the control variable under distributional stress, we evaluate the model on clean and perturbed test sets across both domains. Perturbations include token-level noise, sequence truncation, and sequence shuffling, each applied at multiple severity levels to simulate lexical corruption, information loss, and disruption of sequential structure, respectively. We measure both external robustness, via Accuracy, F1, and ROC-AUC, and internal robustness, via the alignment between α and reliability-related signals. This experiment tests the central hypothesis that the proposed priors improve not only nominal predictive performance but also preserve behaviorally coherent and interpretable control dynamics under input perturbations. Table 2 summarizes robustness under distributional perturbations. In Domain 1, the model remains stable under moderate corruption, with only modest degradation under token noise and truncation, while sequence shuffling induces the strongest performance drop. In Domain 2, all perturbations produce larger declines across predictive metrics, reflecting greater distributional complexity and stronger dependence on sequence structure. Despite these degradations, the behavioral alignment of α remains comparatively stable across perturbation settings in both domains. In Domain 1, α–reliability correlation remains above 0.79 across all conditions; in Domain 2, it remains above 0.78 even under severe perturbations. These results suggest that the proposed priors do not make the model invariant to arbitrary corruption, but preserve the functional coherence of the control mechanism under stress. Overall, the robustness gains are behavioral rather than absolute: while predictive performance degrades under severe perturbations, the internal control dynamics remain aligned with reliability-related signals .

## 5.4 Ablation Study

To isolate the contribution of each behavioral constraint, we conduct an ablation study over both domains. Specifically, we compare the full model against variants that remove either the effect prior or the monotonicity prior, as well as a baseline without function-space priors. This experiment evaluates whether performance gains arise primarily from enforcing functional relevance, directional consistency, or their combination. Models are compared using predictive performance (Accuracy, F1, ROC-AUC) and behavioral indicators related to α stability and alignment. Table 3 summarizes the ablation results across both domains. In Domain 1, all variants achieve similar predictive performance, indicating that the task is relatively easy and less sensitive to behavioral constraints. The Effect Only variant achieves the highest ROC-AUC and strongest α–reliability alignment, suggesting that preserving functional relevance alone is sufficient in simpler settings. In Domain 2, the contribution of each prior becomes more pronounced. The Full Model achieves the highest Accuracy (0.853), F1-score (0.852), and ROC-AUC (0.916), indicating that combining both priors yields the best predictive performance under greater distributional complexity. The No Priors variant exhibits clear collapse behavior, with near-zero α variability and a strongly negative α–reliability correlation, indicating loss of meaningful control dynamics. The ablations further suggest complementary roles for the two priors. The Effect Prior substantially improves α–reliability alignment, indicating its role in preserving functional relevance, while the Monotonicity Prior improves predictive performance but provides weaker behavioral alignment when used in isolation. Together, the two constraints provide the most balanced trade-off between predictive accuracy and stable internal control behavior.

Table 2. Ablation study across domains

| **Domain** | **Variant** | **Accuracy ↑** | **F1 ↑** | **ROC-AUC ↑** | **α–Rel Corr ↑** | **α Std** |
|---|---|---|---|---|---|---|
| D1 | No Priors | 0.948 | 0.948 | 0.984 | 0.660 | 0.285 |
| D1 | Effect Only | 0.953 | 0.953 | 0.986 | 0.902 | 0.298 |
| D1 | Monotonicity Only | 0.952 | 0.953 | 0.981 | 0.817 | 0.290 |
| D1 | Full Model | 0.951 | 0.952 | 0.982 | 0.806 | 0.307 |
| D2 | No Priors | 0.819 | 0.821 | 0.900 | -1.000 | 0.019 |
| D2 | Effect Only | 0.847 | 0.846 | 0.908 | 0.923 | 0.300 |
| D2 | Monotonicity Only | 0.849 | 0.846 | 0.907 | 0.310 | 0.297 |
| D2 | Full Model | 0.853 | 0.852 | 0.916 | 0.867 | 0.300 |

Table 3 shows that the proposed function-space priors improve predictive performance and reduce uncertainty, particularly in the more complex Domain 2. In D2, the model with priors increases test accuracy from 0.820 to 0.856, improves precision from 0.810 to 0.870, raises ROC-AUC from 0.900 to 0.920, and slightly reduces overall model uncertainty. In Domain 1, where performance is already saturated, the gains are more modest, with small improvements in accuracy, loss, recall, and uncertainty. Overall, these results suggest that the behavioral priors are most beneficial under higher distributional complexity, where preserving the functional role of the reliance parameter α contributes to both stronger prediction and more stable uncertainty estimates.

Table 3. Predictive performance and uncertainty metrics with and without function-space priors on the reliance parameter α.

| **Performance Metric** | **D2- No Priors** | **D2- With Priors** | **D1- No Priors** | **D1- With Priors** |
|---|---|---|---|---|
| Test Accuracy | 0.820 | 0.856 | 0.948 | 0.951 |
| Test Loss | 0.0102 | 0.0090 | 0.0072 | 0.0070 |
| Precision | 0.810 | 0.870 | 0.940 | 0.930 |
| Recall | 0.840 | 0.840 | 0.960 | 0.970 |
| F1 Score | 0.820 | 0.850 | 0.950 | 0.950 |
| ROC-AUC | 0.900 | 0.920 | 0.980 | 0.980 |
| Overall model uncertainty | 0.011 | 0.010 | 0.028 | 0.023 |

Table 4 evaluates the robustness of the proposed model under distributional perturbations, including token noise, sequence truncation, and sequence shuffling. As expected, predictive performance decreases as perturbation severity increases, with stronger degradation in Domain 2, suggesting that this domain is more sensitive to distributional stress and sequence disruption. However, the α–reliability correlation remains consistently high across all perturbation settings, staying above 0.79 in Domain 1 and above 0.78 in Domain 2. This indicates that although external predictive performance declines under severe input corruption, the reliance parameter α continues to preserve its behavioral alignment with reliability-related signals. Overall, the results suggest that the proposed function-space priors improve internal robustness by maintaining coherent control dynamics under distributional shift.

Table 4. Robustness Evaluation under Input Perturbations

| Domain | Condition | Accuracy ↑ | F1 ↑ | ROC-AUC ↑ | α–Rel Corr ↑ |
|---|---|---|---|---|---|
| D1 | Clean | 0.951 | 0.952 | 0.984 | 0.808 |
| D1 | Noise 10% | 0.944 | 0.945 | 0.980 | 0.803 |
| D1 | Noise 20% | 0.932 | 0.934 | 0.973 | 0.799 |
| D1 | Trunc 10% | 0.947 | 0.949 | 0.982 | 0.801 |
| D1 | Trunc 25% | 0.940 | 0.942 | 0.978 | 0.796 |
| D1 | Shuffle 10% | 0.912 | 0.905 | 0.955 | 0.791 |
| D1 | Shuffle 25% | 0.872 | 0.863 | 0.910 | 0.793 |
| D2 | Clean | 0.856 | 0.854 | 0.921 | 0.867 |
| D2 | Noise 10% | 0.790 | 0.781 | 0.880 | 0.839 |
| D2 | Noise 20% | 0.733 | 0.719 | 0.825 | 0.810 |
| D2 | Trunc 10% | 0.801 | 0.792 | 0.891 | 0.851 |
| D2 | Trunc 25% | 0.767 | 0.762 | 0.853 | 0.823 |
| D2 | Shuffle 10% | 0.812 | 0.804 | 0.900 | 0.842 |
| D2 | Shuffle 25% | 0.779 | 0.769 | 0.861 | 0.782 |

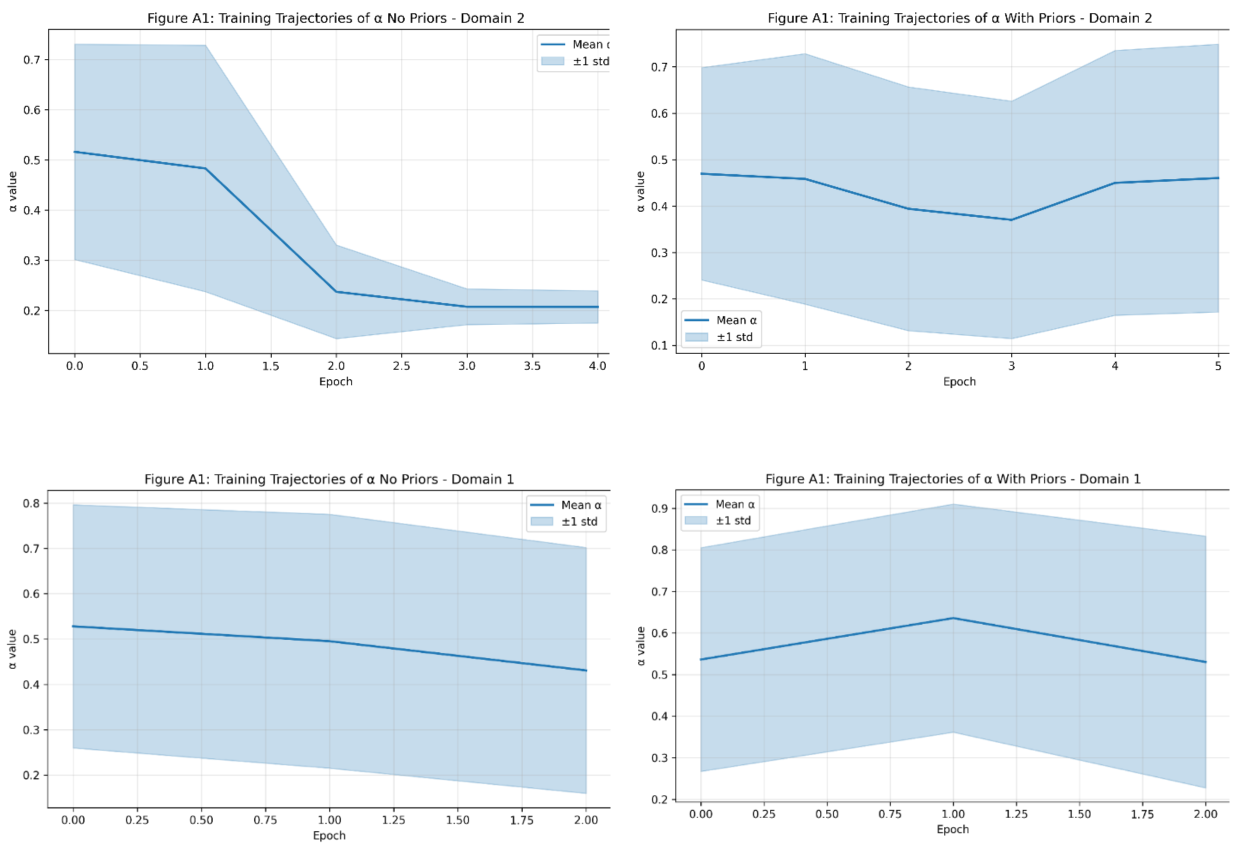


Figure 1. Effect of Function-Space Priors on α Dynamics in Domain 1 and Domain 2

Figure 1 illustrates that function-space priors stabilize the training dynamics of the reliance parameter α. In both domains, α decreases under the no-prior condition, with a sharper collapse in Domain 2, whereas the use of priors preserves higher and more variable α trajectories across epochs. These results indicate that the proposed priors help prevent low-reliance collapse and maintain the functional role of α during optimization.

The empirical results suggest that the proposed function-space priors do more than improve predictive performance or robustness: they alter the optimization dynamics governing the emergence of internal control behavior. Across domains, the priors systematically reshape the operating regime of the reliance signal α, increasing its alignment with retrieval-related evidence and preventing collapse into weakly-informative or behaviorally inconsistent states. This indicates that auxiliary control variables are not inherently meaningful simply because they are architecturally present; rather, their functional role must be actively preserved during optimization. This observation points to a broader phenomenon in modern neural systems: unconstrained optimization may achieve strong nominal performance while silently degrading internal mechanisms. In this view, collapse is not limited to posterior collapse in latent-variable models or entropy collapse in attention, but may reflect a more general tendency of optimization to suppress or trivialize auxiliary pathways when their contribution is not explicitly incentivized. The results here provide empirical evidence of such a mechanism in reliance-based arbitration. Importantly, the proposed priors operate in function space, constraining behavioral sensitivity rather than parameter values. This distinction may explain why classical parameter-space regularization proves insufficient under complex settings. While weight-space regularization penalizes capacity globally, function-space constraints target the local behavioral geometry of decision-making, shaping how the model responds rather than merely limiting what it can represent. The domain-dependent effects further suggest that meaningful internal behavior is most critical under distributional complexity. In saturated tasks, behavioral regularization primarily improves stability and interpretability; under higher uncertainty, the same constraints translate into measurable gains in predictive performance and robustness. This implies that internal behavioral coherence may function as a latent source of generalization under stress. More broadly, these findings motivate a shift in perspective: auxiliary control variables should not be viewed merely as architectural artifacts or post-hoc interpretability handles, but as dynamical mechanisms whose semantics can drift during training. Preserving their functional meaning may therefore be a necessary design principle in reliability-aware, retrieval-augmented, and memory-driven learning systems. Taken together, this work suggests that function-space behavioral priors provide not only an inductive bias for interpretability, but a mechanism for preserving semantically coherent internal computation under optimization pressure.

## 6. Conclusion

We introduced lightweight function-space behavioral priors to preserve the functional role of an auxiliary reliance variable under optimization and distributional stress. By explicitly constraining the local sensitivity and directional consistency of the decision function with respect to $\alpha$, the proposed method reshapes internal control dynamics without imposing architectural restrictions or significant computational overhead. Across multiple experiments, the priors consistently improved the alignment of $\alpha$ with retrieval-related signals, stabilized its training dynamics, and yielded gains in robustness, uncertainty reduction, and predictive performance under more complex conditions. These results suggest that unconstrained optimization can silently degrade auxiliary control mechanisms even when nominal task performance remains high, and that explicitly regularizing behavior in function space provides a principled inductive bias for preserving semantically coherent internal computation. More broadly, this work supports the view that internal behavioral coherence may itself be a source of robustness and generalization in reliability-aware learning systems.

## References


Bornacelly Castaneda, M., BayesIntuit: A neural framework for intuition-based reasoning, *Communications in Computer and Information Science (CCIS)*, vol. 2557, Springer Nature, 2025. https://doi.org/10.1007/978-3-031-98235-4_9

Daniels, H. and Velikova, M., Monotone and partially monotone neural networks, *IEEE Transactions on Neural Networks*, vol. 21, no. 6, pp. 906-917, 2010. https://doi.org/10.1109/TNN.2010.2041833

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q., On calibration of modern neural networks, *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321-1330, 2017.

Krogh, A. and Hertz, J. A., A simple weight decay can improve generalization, *Advances in Neural Information Processing Systems*, vol. 4, 1991.

Lucas, J., Tucker, G., Grosse, R., and Norouzi, M., Understanding posterior collapse in generative latent variable models, *ICLR 2019 Workshop on Deep Generative Models for Highly Structured Data*, 2019.

MacKay, D. J. C., The evidence framework applied to classification networks, *Neural Computation*, vol. 4, no. 5, pp. 720-736, 1992.

Neal, R. M., Bayesian Learning for Neural Networks, *Springer*, 1996.

Rudner, T. G. J., Fortuin, V., Gal, Y., and Mandt, S., Function-space regularization in neural networks: A probabilistic perspective, *Proceedings of the 40th International Conference on Machine Learning*, 2023. https://doi.org/10.48550/arXiv.2312.17162

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R., Dropout: A simple way to prevent neural networks from overfitting, *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929-1958, 2014.

Zhai, S., Likhomanenko, T., Littwin, E., Busbridge, D., Ramapuram, J., Zhang, Y., Gu, J., and Susskind, J., Stabilizing transformer training by preventing attention entropy collapse, *arXiv preprint*, 2023. https://doi.org/10.48550/arXiv.2303.06296

## Biography

**Mayra Bornacelly** is a Ph.D. student in Industrial Engineering at the University of Central Florida, USA, where she also serves as a Graduate Teaching Associate. Her research focuses on the reliability and fairness of machine learning systems under distributional and temporal shift, with an emphasis on representation-level diagnostics grounded in optimization geometry. She integrates perspectives from uncertainty quantification, interpretability, and human factors to study how learned systems sustain decisions under changing operating conditions. Her email address is mayra.bornacellycastaneda@ucf.edu.