Limits of Monitoring

Kieran Tran

December 30, 2025

Executive Summary

What problem am I trying to solve?

A central challenge in AI safety is whether we can reliably *monitor a trained model for dangerous internal states*, such as deceptive intent, hidden objectives, or other safety-relevant latent variables. Many proposed approaches implicitly assume that sufficiently powerful monitors such as neural probes applied to internal activations, can detect such states if they exist.

This project asks a more fundamental question:

> *When is reliable monitoring possible in principle, and when is it impossible regardless of monitor capacity?*

I formalize monitoring as a hypothesis testing problem over internal representations. To study this question in a controlled yet dynamic setting, I use **state-space models (SSMs)** and instantiate a latent safety variable Z as a **binary internal mechanism** governing the model's state evolution. A monitor observes an internal activation H derived from the SSM's latent state and attempts to infer Z. This framing allows monitoring to be formalized as a hypothesis testing problem over internal representations, where feasibility is determined by the statistical distinguishability of the induced distributions $P(H \mid Z = 0)$ and $P(H \mid Z = 1)$.

Code availability. All experiments were implemented in Python using PyTorch. The full codebase, including model definitions, training scripts, and configuration files required to reproduce the results, is available at: `https://github.com/Isomorphic-07/State-Space-Models-Monitoring/blob/main/monitoring_ssm.py`.

High-level takeaways

- Monitoring accuracy has a tight **information-theoretic upper bound** determined by the total variation (TV) distance between the induced internal-state distributions, not by monitor capacity.
- Increasing monitor capacity exhibits **diminishing returns**: accuracy plateaus once the Bayes-optimal decision rule is approximated.
- Sparse autoencoders (SAEs) reveal a **non-trivial tradeoff**: moderate sparsity can improve finite-sample monitoring performance via denoising, while stronger sparsity destroys safety-relevant information and reduces monitorability.
- Monitoring failure can arise from **representational unfaithfulness**, not merely from insufficiently expressive monitors.

Together, these results show that successful monitoring depends critically on how safety-relevant information is encoded in internal representations, not just on the sophistication of the monitoring method.

Key experiments and results

1. Capacity sweep: existence of an accuracy ceiling I instantiated the latent safety variable Z as a switch between two state-space model (SSM) transition mechanisms and trained monitors on the resulting hidden states. As monitor capacity increased from linear probes to large MLPs, validation accuracy improved initially but then *plateaued sharply.*

This plateau empirically demonstrates the theoretical bound: once the monitor is expressive enough, further capacity cannot overcome overlap between $P(H \mid Z = 0)$ and $P(H \mid Z = 1)$. Monitoring becomes limited by the representation channel itself rather than by the expressiveness of the probe.

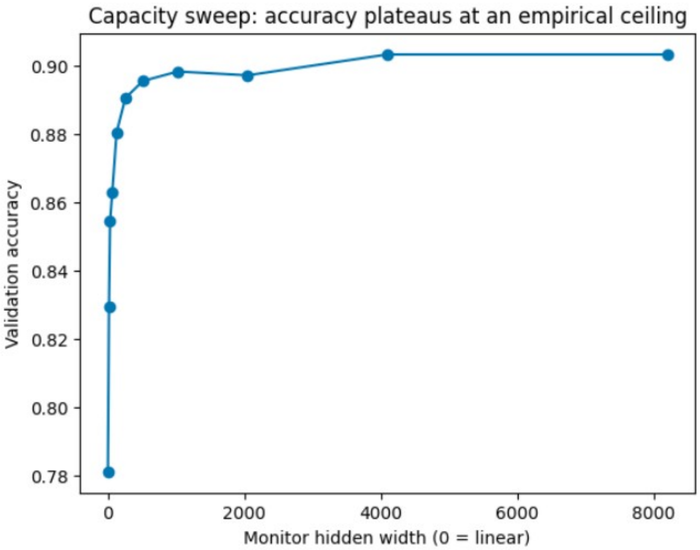

Figure 1: Capacity Sweep, increasing hidden width of MLP Monitor

2. Sanity check: accuracy equals TV witness For each trained monitor M, we define its decision region $A = \{h \in \mathbb{R}^d : M(h) = 1\}$, i.e. the set of internal activations h for which the monitor predicts the dangerous latent state $Z = 1$. Let $P_1(A)$ and $P_0(A)$ denote the probabilities that the internal representation H falls in A under the two latent mechanisms $Z = 1$ and $Z = 0$, respectively, and let $\pi = P(Z = 0)$ (here, we let $\pi = 0.3$ to study a moderately imbalanced but non-degenerate regime in which Bayes-optimal accuracy is neither trivial nor saturated) be the prior probability of the benign mechanism. Theory predicts that the monitoring accuracy should equal $\pi + \big((1 - \pi)P_1(A) - \pi P_0(A)\big)$, which is precisely the total-variation witness associated with the monitor's decision rule. Empirically, we find that the observed validation accuracy lies almost exactly on the identity line when plotted against this quantity, confirming that trained monitors correspond directly to measurable decision regions in the theory and that their performance is governed by total variation separation rather than monitor capacity.

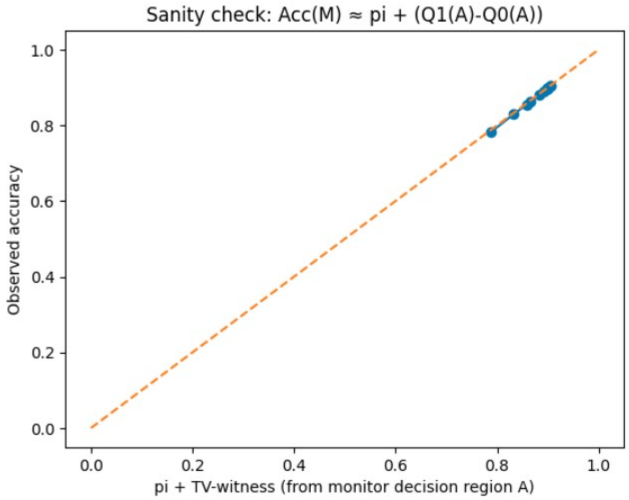

Figure 2: Sanity Check

3. Sparsity sweep: compression can reduce monitorability To study representational faithfulness, I inserted a sparse autoencoder between the internal state H and the monitor, yielding compressed codes U. Varying the sparsity penalty revealed **non-monotonic behavior**:

- At low to moderate sparsity, monitoring accuracy on U sometimes exceeded that of a reference monitor trained directly on H, due to denoising and improved estimator efficiency.
- At higher sparsity, reconstruction error increased and monitoring accuracy dropped below baseline, indicating genuine information loss.

This experiment cleanly separates *estimation limits* from *information-theoretic limits*, showing that compression can both help and harm monitoring depending on the regime.

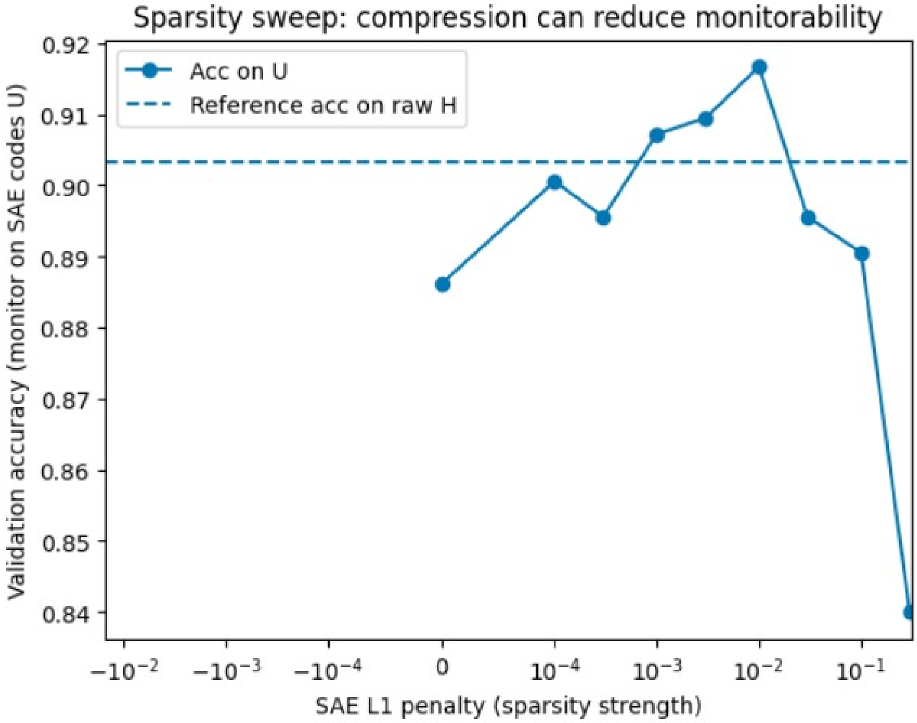

Figure 3: Sparsity Sweep

Why this matters

These results caution against treating monitoring as purely an engineering problem. Even perfect monitors cannot succeed if internal representations are not faithful to safety-relevant variables. Conversely, failures in practice may stem from estimation issues rather than fundamental impossibility. Overall, this work hopes to provide a theoretical and empirical framework for understanding **when monitoring internal states is feasible, and when it is fundamentally constrained**.

1 Theory

First, we define the latent safety variable:

$$Z \in \{0,1\}$$

Which is a binary random variable representing a latent safety-relevant property of a trained model. Here, $Z = 1$ represents the presence of a dangerous interval state (i.e. deceptive intent, hidden objective etc.), and $Z = 0$ the absence of such a state. Often, this variable is not observable.

We now also define an abstraction for the internal representation, let:

$$H \in \mathbb{R}^d$$

be a random vector representing the internal state of the model accessible to a monitor. Concretely, H may correspond to a hidden activation vector like a residual stream or MLP activation at a fixed layer and token position.

We will also assume that the distribution of internal states depends on the latent variable Z. Specifically, for each $z \in \{0,1\}$, let

$$P(H|Z = z)$$

denote the conditional distribution of the internal representation H given $Z = z$. Additionally, we write:

$$P_0 = P(H|Z = 0)$$

$$P_1 = P(H|Z = 1)$$

We will also define the notion of a monitor, which is any measurable function:

$$M : \mathbb{R}^d \to \{0,1\}$$

Which, given an observed internal representation H, outputs a prediction $M(H)$ of the latent variable Z. Additionally, the accuracy of a monitor is defined as:

$$\mathrm{Acc}(M) = \mathbb{P}[M(H) = Z]$$

For generality, we assume a binary Bayes setting with the following priors:

$$\mathbb{P}(Z = 0) = \pi$$

$$\mathbb{P}(Z = 1) = 1 - \pi$$

For intuition, we will also say that the monitor M is **nontrivial** if:

$$\mathrm{Acc}(M) \geq \max\{1 - \pi, \pi\} + \varepsilon$$

for some $\varepsilon > 0$. Essentially, this is saying that the monitor can indicate the true latent safety variable more accurate than random (provided priors), where the accuracy of the monitor is strictly better than chance. To evaluate how "different" 2 probability distributions are, we will specifically consider the Total Variation distance:

$$D_{TV}(P, Q) = \frac{1}{2} \sum_{x \in \Omega} |P(x) - Q(x)| = \sup_{A \subseteq \Omega} |P(A) - Q(A)|$$

Where P, Q are probability distributions. Notice that this notion of a measure of distance between 2 probability distributions is symmetric and bounded in range $[0,1]$, which are important characteristics in analyzing the limits of monitoring. This is contrary to other measures such as **KL divergence** which does not directly upper-bound classification accuracy and is not symmetric. We now provide a proposition:

Proposition 1 (Fundamental Limit of Monitoring)**.** *Let* $Z \in \{0,1\}$ *be a binary random variable with* $\mathbb{P}(Z = 0) = \pi$, $\mathbb{P}(Z = 1) = 1 - \pi$. *Let* $H \in \mathbb{R}^d$ *be an internal representation with conditional distributions*

$$P_0 := P(H \mid Z = 0), \qquad P_1 := P(H \mid Z = 1).$$

Then for any (measurable) monitor $M : \mathbb{R}^d \to \{0,1\}$*,*

$$\mathbb{P}[M(H) = Z] \;\leq\; \pi + TV((1 - \pi)P_1, \pi P_0),$$

where the total variation distance is defined by

$$\mathrm{TV}(P_0, P_1) := \sup_{A \subseteq \mathbb{R}^d} |P_0(A) - P_1(A)| .$$

Moreover, this bound is tight:

$$\sup_M \mathbb{P}[M(H) = Z] \ = \ \pi + TV((1-\pi)P_1, \pi P_0).$$

Proof. We first define the idea of Accuracy. We consider the following set:

$$A = \{h \in \mathbb{R}^d ; M(h) = 1\}$$

Then, $M(H) = 1$ iff $H \in A$ and $M(H) = 0$ iff $H \notin A$. So accuracy is:

$$P(M(H) = Z) = P(Z = 1, H \in A) + P(Z = 0, H \notin A)$$

We can further simplify this via Bayes:

$$P(Z = 1, H \in A) = P(Z = 1) \cdot P(H \in A|Z = 1) = (1 - \pi) \cdot P(H \in A|Z = 1) = (1 - \pi)P_1(A)$$

$$P(Z = 0, H \notin A) = P(Z = 0) \cdot P(H \notin A|Z = 0) = \pi \cdot P(H \notin A|Z = 0) = \pi(1 - P_0(A))$$

$$\therefore P(M(H) = Z) = (1 - \pi)P_1(A) + \pi(1 - P_0(A)) = \pi + (1 - \pi)P_1(A) - \pi P_0(A)$$

Let us now bound this accuracy via the supremum:

$$\sup_M \mathrm{Acc}(M) = \pi + \sup_A ((1 - \pi)P_1(A) - \pi P_0(A))$$

Let:

$$Q_1(A) = (1 - \pi)P_1(A)$$

$$Q_0(A) = \pi P_0(A)$$

We will also denote:

$$\Delta(A) = Q_1(A) - Q_0(A)$$

Let Ω be the entire sample space, so since Q_0 and Q_1 are probability measures:

$$Q_1(\Omega) = Q_0(\Omega) = 1$$

$$\therefore \Delta(\Omega) = 0$$

For any measurable A

$$\Delta(A^c) = Q_1(A^c) - Q_0(A^c) = Q_1(\Omega) - Q_1(A) - Q_0(\Omega) + Q_0(A) = -\Delta(A)$$

So by symmetry, we can see that every negative value of $\Delta(A)$ appears as a positive value via the complement! Hence:

$$|\Delta(A)| = \max\{\Delta(A), \Delta(A^c)\}$$

Note, since A is a measurable set:

$$\sup_A |\Delta(A)| = \sup_A \max\{\Delta(A), \Delta(A^c)\} = \sup_A \Delta(A)$$

By the definition of the Total Variation distance:

$$\sup_A \Delta(A) = TV(Q_1, Q_0) = TV((1 - \pi)P_1, \pi P_0)$$

Thus,

$$\sup_M \mathrm{Acc}(M) = \pi + TV((1 - \pi)P_1, \pi P_0)$$

□

The crucial idea that we observe here is that the accuracy of monitoring has a tight upper bound, which inherently tells us that mathematically, there are in built limitations to the monitor's accuracy. We now revisit the idea of non-triviality of monitors, for some $\varepsilon > 0$:

$$\max\{1-\pi, \pi\} + \varepsilon \leq \mathrm{Acc}(M) \leq \pi + TV((1-\pi)P_1, \pi P_0)$$

$$\therefore TV((1-\pi)P_1, \pi P_0) \geq \max\{1-\pi, \pi\} - \pi + \varepsilon$$

The inequality above can be directly interpreted via the concept of faithfulness. Recall that the total variation distance characterizes the maximal distinguishability between 2 measures over all measurable events. So given that the monitor is non-trivial with accuracy exceeding $\max\{1-\pi, \pi\} + \varepsilon$ requires that the observational measures $(1-\pi)P_1$ and πP_0 be separated by a non-vanishing margin in total variation. When this condition is not satisfied, no monitoring procedure regardless of its computational power can reliably distinguish the underlying mechanisms. Ultimately, limitations on monitoring accuracy comes from an intrinsic lack of faithfulness in the observation channel, rather than from deficiencies of the monitor itself. Faithfulness here can be defined formally: We say an observation channel is ε-faithful if:

$$\mathrm{TV}((1-\pi)P_1, \pi P_0) \geq \epsilon$$

Faithfulness here does not require that the latent variable Z be linearly decodable or sparse, only that it induces separable distributions over internal states.

2 Experimentation: Mechanistic Uncertainty in State Space Models

To experimentally probe the theoretical limits established, we instantiate the latent safety variable Z within a state space model. where the model's internal latent state provides a concrete and temporally persistent representation accessible to intervention. We use state-space models because they expose a persistent latent state through which all long-horizon behavior must flow, allowing a clean definition of the observation channel H. While our results are not architecture-specific, SSMs provide a controlled substrate in which latent mechanisms and information flow can be precisely analyzed. In contrast to purely feed-forward or attention-based architectures, SSMs explicitly maintain a latent state s_t that evolves according to a transition mechanism and mediates all future outputs. This makes the internal state a natural candidate for the abstract variable H in our framework. Here, we construct neural SSM's which combine structured state evolution with learned, input-dependent dynamics. In these models, the latent state update:

$$s_{t+1} = f_\theta(s_t, x_t)$$

defines a persistent internal channel through which information and potentially safety relevant intent must flow. By construction of SSM's any latent property that influences long-horizon behaviour must be encoded in the distribution of the latent states s_t, allowing us to operationalize the conditional distributions $P(H|Z=0)$ and $P(H|Z=1)$ as distributions over internal states under different latent regimes. The experiment will treat the presence or absence of a dangerous internal mechanism as the latent variable Z, while monitors are restricted to observing internal activations derived from the SSM state.

Then, we empirically estimate the total variation distance between the induced state distributions corresponding to different values of Z. This would directly test whether the observation channel provided by the SSM latent state is faithful in the sense described above.

This stepup allows us to distinguish failures of monitoring due to insufficient separability of internal representations from failures due to monitor capacity or training. When total variation distance between the induced state distributions is small, our theory predicts that no monitor regardless of complexity, can reliably detect the latent mechanism. SSMs serve not just as a modeling choice, but also as a controlled experimental substrate for validating fundamental limits on internal state monitoring.

This framework also enables targeted interventions on the state transition dynamics themselves, allowing us to study how latent safety variables propagate, decay or become entangled over time.

Let us lay the mathematical foundations of this, then present the code implementation. We define the following SSM:

$$s_{t+1} = f_\theta(s_t, x_t)$$

$$h_t = g_\theta(s_t) \in \mathbb{R}^d$$

Here, our monitor observes $H = h_t$. We will define the latent safety variable as a mechanism switch:

$$Z = 0 : \text{benign dynamics: } f_\theta^{(0)}$$

$$Z = 1 : \text{dangerous dynamics: } f_\theta^{(1)}$$

So, the latent state update is:

$$s_{t+1} = \begin{cases} f_\theta^{(0)}(s_t, x_t), & Z = 0 \\ f_\theta^{(1)}(s_t, x_t), & Z = 1 \end{cases}$$

In the experiments, we instantiate Z as a switch between 2 SSM transition mechanisms and treat the induced hidden-state distribution $P(H|Z = z)$ as the observation channel. Monitors trained on hidden states act as empirical witnesses for total variation separation, directly testing the theory's predicted limits on detectability.

For the SSM we design it via:

$$\tilde{s}_{t+1} = f_\theta^{(Z)}(s_t, x_t)$$

We can think of this as the proposed update using the usual state transition. However, we also introduce an input-dependent gate:

$$g_t = \sigma(W_g x_t) \in (0, 1)^d$$

Where W_g is a learned matrix, $\sigma(\cdot)$ is sigmoid applied elementwise, g_t has the same dimension as state $s_t \in \mathbb{R}^d$. Now we define the gated update equation:

$$s_{t+1} = g_t \cdot \tilde{s}_{t+1} + (1 - g_t) \cdot s_t$$

This allows us to maintain long-range information without constantly drifting which is very important due to the SSM transition mechanisms being dependent on Z. We also require a continuous way to make the two mechanisms more/less different. So we implement a separability knob α:

$$f_\alpha^{(0)} = f^{(0)}$$

$$f_\alpha^{(1)} = (1 - \alpha) f^{(0)} + \alpha f^{(1)},\ \alpha \in [0, 1]$$

Hence, $\alpha = 0$ means that the mechanisms are identical implying that $TV = 0$ and so detection becomes impossible. To ensure the experiment is focused on detectability only to mechanism difference, the following are kept the same:

Same initial state distribution

$$s_0 \sim \rho \text{ (often just zeros)}$$

Same input distribution (i.e. Gaussian sequences)

$$x_{1:T} \sim \mathcal{D}$$

Now we also need to define the monitor dynamics. We can implement a linear monitor which assumes separability is captured by some hyperplane:

$$M(h) = \mathbf{1}[w^\top h + b \geq 0]$$

Additionally, we could also implement an MLP for more complex decision boundaries.

3 Discussion of Results

We now discuss the results of the experiment. It should be noted here that we initialize $\alpha = 0.1$ and $\pi = 0.3$. We will first experimentally show the existence of the upper bound on the Bayes optimal monitoring accuracy:

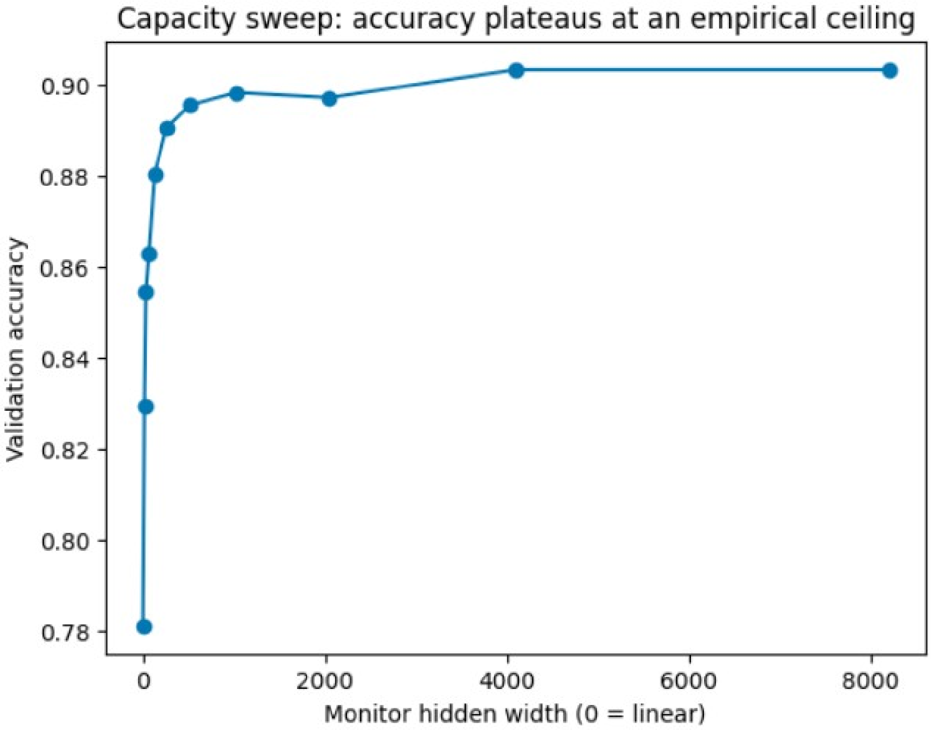

Figure 4: Capacity Sweep, increasing hidden width of MLP Monitor

Crucially, we show that monitors are bounded not by computational power, but by an information ceiling where monitor capacity is not the bottleneck past some point but rather the representation channel is. In the experiment we test the following hidden layer widths of the MLP monitor:

$$h = [0, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192]$$

Here, the accuracy plateaus to a value of around 0.903. As monitor width increases, validation accuracy rises rapidly at small capacities but then saturates, exhibiting a clear plateau. This behavior is consistent with the theoretical result that the optimal achievable accuracy is upper bounded by the total variation separation between induced distributions $P(H|Z = 0)$ and $P(H|Z = 1)$.

To verify we are indeed correctly observing this Bayes monitoring accuracy, we will implement a sanity check:

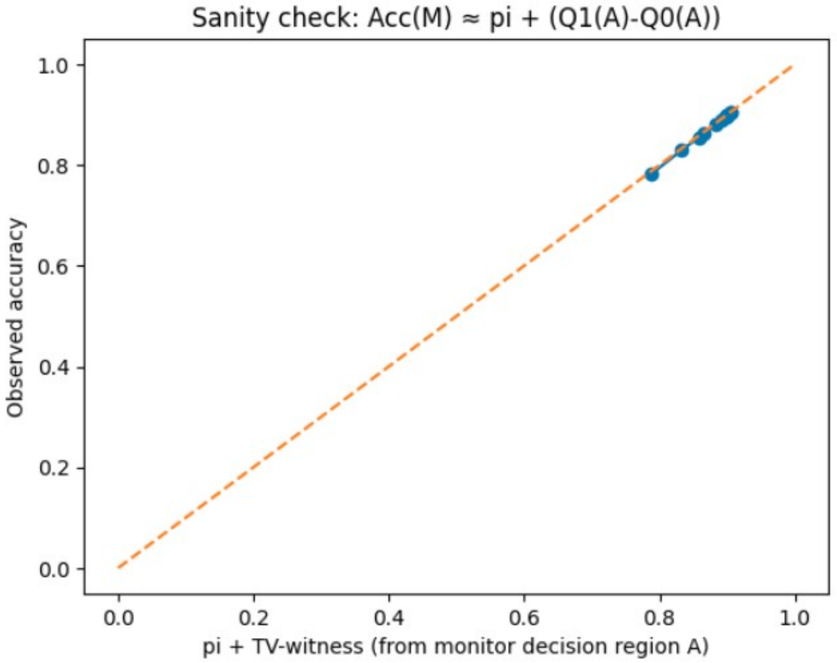

Figure 5: Sanity Check

The figure above verifies the theoretical decomposition of monitor accuracy. For each trained monitor, the observed validation accuracy matches $\pi + (Q_1(A) - Q_0(A))$, where A is the monitor's decision region. This confirms that empirical monitors correspond directly to the abstract measurable sets in the proof, and that the TV distance accurately captures each monitor's contribution toward the optimal bound.

Although post-processing cannot increase total variation in principle, finite-sample training can benefit from representational denoising, leading to apparent improvements in monitoring accuracy.

To study how representational compression affects monitorability, we introduce a sparse autoencoder (SAE) between the internal representation H and the monitor, yielding a compressed code $U = \phi(H)$. This links to our focus on when does a representation permit monitoring at all, and when does it destroy monitorability. From an information-theoretic perspective, post-processing cannot increase the total-variation separation between the conditional distributions $P(H \mid Z)$; consequently, the optimal achievable monitoring accuracy on U is upper bounded by that on H. However, in finite-sample settings, sparsity can substantially influence how closely a trained monitor approaches this bound. Experimentally for the SAE section, we fix the hidden layer width of the MLP monitor to be $h = 4096$ as this value yields the previous validation accuracy of 0.903

Empirically, we observe a non-monotonic relationship between sparsity strength and monitoring performance. At low to moderate sparsity levels, monitoring accuracy on the SAE codes U can exceed that of a reference monitor trained directly on the raw representation H, despite near-perfect reconstruction. This improvement reflects an estimator-level effect rather than an increase in underlying information where the SAE acts as a denoising reparameterization that suppresses nuisance variation and aligns safety-relevant directions with a smaller number of coordinates, thereby improving the signal-to-noise ratio available to the monitor. We also note that since deterministic post-processing cannot increase total variation distance (by the data-processing inequality), any improvement in monitoring accuracy after SAE compression must arise from estimator-level effects rather than increased information about Z. In this regime, sparsity facilitates more efficient approximation of the Bayes-optimal decision rule under limited data and optimization. This can be paralleled to the analogy of PCA, where we remove signal directions corresponding to low variance, and so removing too many signals can remove important information (similar in this context in increasing sparsity).

As sparsity is increased further, reconstruction error rises and the fraction of active features decreases sharply. In this high-sparsity regime, monitoring accuracy degrades below the raw-representation baseline, indicating genuine information loss. This transition marks the point at which compression ceases to be a benign reparameterization and instead reduces the separability of the induced distributions $P(U \mid Z = 0)$ and $P(U \mid Z = 1)$ (think of this as decreasing the TV distance), thereby tightening the theoretical upper bound on monitor performance.

We use sparse autoencoders precisely because they provide a controlled and interpretable way to probe representation faithfulness. By tuning sparsity, we can smoothly interpolate between near-identity transformations and aggressive compression, and empirically distinguish failures of monitoring due to estimator inefficiency from failures due to information-theoretic limitations. This distinction is critical for understanding when monitoring can be improved through better tools, and when it is fundamentally constrained by how safety-relevant information is encoded.

Note in the figure below, we achieve a validation accuracy on the monitor of 0.917 at an L1 penalty value of 10^{-2}.

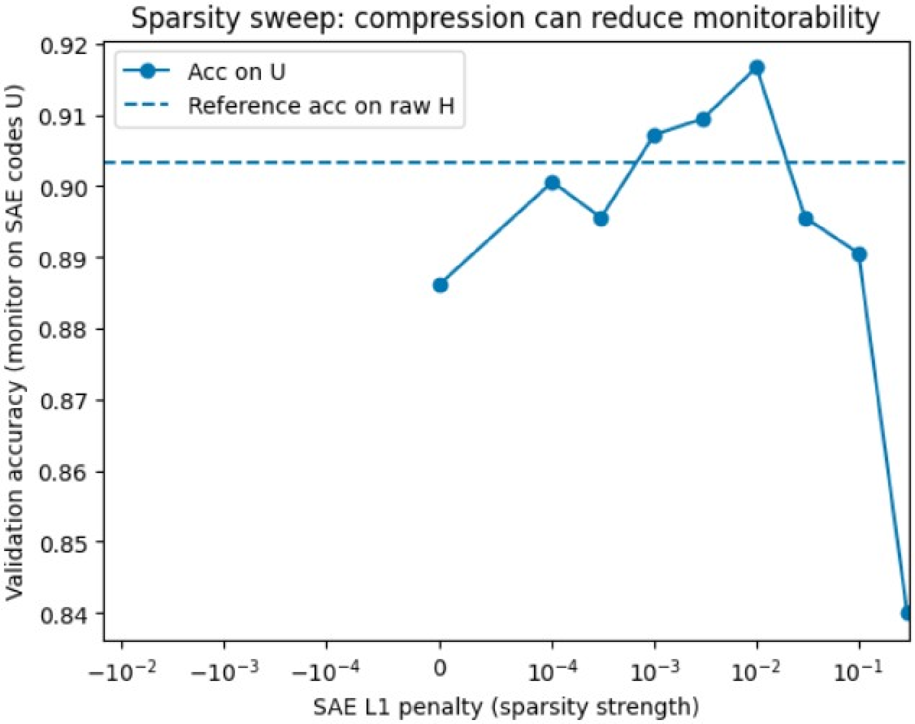

Figure 6: Sparsity Sweep

SAE L_1	Acc on U	TV witness	Recon MSE	Frac. active
0.0	0.886	0.589	0.0000	99.1%
1e-4	0.901	0.603	0.0000	98.9%
3e-4	0.896	0.599	0.0000	99.3%
1e-3	0.907	0.609	0.0001	98.5%
3e-3	0.909	0.612	0.0001	95.9%
1e-2	0.917	0.618	0.0003	82.2%
3e-2	0.896	0.598	0.0003	63.1%
1e-1	0.891	0.592	0.0004	47.9%
3e-1	0.840	0.542	0.0017	32.0%

Table 1: Effect of SAE sparsity on monitoring performance. Accuracy and TV witness are reported for an MLP-4096 monitor trained on SAE codes U. Reference accuracy on raw representations is 0.903 (TV witness = 0.605), corresponding to a Bayes-optimal estimate $\pi + \text{TV} \approx 0.905$. Frac. active tells us the percentage of latent state units that are active by measuring the magnitude of the latent code

We also observe a mild non-monotonicity in the empirical TV witness as a function of sparsity, with the witness peaking at approximately 0.618 for an L_1 penalty of 10^{-2}. Since the SAE implements a deterministic post-processing of the internal representation H, the data-processing inequality implies that the true total variation separation between $P(U \mid Z = 0)$ and $P(U \mid Z = 1)$ cannot exceed that of the original representations. We therefore interpret this increase not as a fundamental gain in information, but as an estimator-level effect. Moderate sparsification improves the conditioning and denoising of the representation, allowing finite-sample monitors to more closely approximate the Bayes-optimal decision boundary. At higher sparsity levels, this benefit is outweighed by genuine information loss, leading to a decline in both the TV witness and monitoring accuracy.

4 Limitations and Future Directions

While this work establishes a tight and operationally meaningful limit on the monitorability of internal states, it is subject to several important limitations that point toward directions for future research.

Snapshot-based monitoring. Our analysis focuses on monitors that observe a single internal representation H at a fixed time. While this abstraction is sufficient to expose fundamental information-theoretic limits, it ignores the temporal structure inherent in dynamical models such as state-space models. In practice, a monitor may have access to a sequence of internal states $\{H_t\}_{t=1}^{T}$, enabling sequential decision rules and hypothesis tests. Extending the framework to sequential monitoring would allow characterization of error exponents and time-dependent detectability, potentially revealing regimes in which latent safety variables become identifiable only through their long-horizon effects on state evolution. This would ultimately change our notion that in the experiments $\pi = P(Z = 0)$ is a fixed deterministic value, when it could be a probabilistic function of time.

Binary latent mechanisms. We model the latent safety variable as a binary switch between two transition mechanisms. This choice enables a clean theoretical characterization but abstracts away the graded, compositional, or multi-factor nature of real safety-relevant properties. An important extension would consider multi-valued or continuous latent variables, leading to multiclass or composite hypothesis testing problems. In such settings, total variation separation between mixtures of internal-state distributions may yield richer notions of partial monitorability and failure modes.

Passive observation. The monitoring framework studied here is purely observational: monitors passively inspect internal representations without intervening on the model's dynamics. In many interpretability and safety settings, however, targeted interventions on internal states are possible. Future work could study interventional monitoring, where perturbations to latent states or transition dynamics are used to amplify separability between safety regimes. This would connect the present framework to causal and counterfactual approaches to mechanistic interpretability.

Architecture-specific instantiation. Our experiments use state-space models as a controlled substrate for studying internal dynamics and latent mechanisms. While SSMs provide explicit, persistent latent states that are well-suited to the theoretical framework, the results are not specific to this architecture. Extending the analysis to transformers, recurrent networks, or hybrid architectures would help clarify how architectural choices influence representational faithfulness and the resulting limits on monitoring.

Finite-sample and optimization effects. Although total variation distance characterizes the optimal achievable monitoring accuracy in the limit of infinite data and ideal optimization, practical monitors operate in finite-sample regimes. Our sparse autoencoder experiments illustrate that representational reparameterizations can substantially affect how closely trained monitors approach the theoretical bound. A deeper theoretical treatment of finite-sample effects, optimization dynamics, and implicit regularization would further clarify when monitoring failures reflect estimator inefficiency versus fundamental information loss.

Overall, these limitations suggest that while monitoring accuracy is fundamentally constrained by representation-level information, important open questions remain regarding how temporal structure, interventions, architectural design, and finite-sample effects shape the practical feasibility of detecting safety-relevant internal states.

5 Conclusion

In this work, we studied the fundamental limits of monitoring internal states by framing monitoring as a hypothesis testing problem over internal representations. Using state-space models as a controlled setting for studying internal dynamics, we showed both theoretically and empirically that monitoring accuracy is bounded by the statistical distinguishability of the induced representation distributions, rather than by the expressiveness of the monitor itself. Our experiments demonstrate that increasing monitor capacity yields diminishing returns, converging to a representation-dependent ceiling, and that representational compression via sparse autoencoders can either improve finite-sample estimation or irreversibly destroy safety-relevant information depending on the regime. Together, these results highlight a critical distinction between estimator limitations and information-theoretic constraints, and emphasize that monitoring failure can arise from unfaithful internal representations rather than inadequate monitoring tools. More broadly, this framework provides a principled lens for evaluating the feasibility of internal monitoring in complex dynamical models, and suggests that progress in AI safety will require not only better monitors, but also representations that reliably expose safety-relevant latent variables.

References

[1] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231(694–706):289–337, 1933.

[2] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory.* Springer, 1986.

[3] Y. Bengio et al. Towards causal representation learning. *Nature Machine Intelligence*, 2023.

[4] A. Ng. Sparse autoencoder. CS294A Lecture Notes, Stanford University, 2011.

[5] A. Wald. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16(2):117–186, 1945.