Hidden Gauges in Learned Representations

D. Tanton

December 2025

Abstract

Learning agents interact with the world through a specific interface of sensors, actuators, and reward mechanisms. In standard training, these interface conditions are often fixed, causing models to implicitly hard-code parameters that should ideally be treated as variable. This paper argues that such parameters function as **gauges**: degrees of freedom that do not change the underlying world state but globally reparameterize the geometry of the agent's latent representation. I introduce a diagnostic toolkit to measure this structure in standard deep learning models, distinguishing hidden gauges from nuisance variables using three coupled metrics: *sensitivity* (responsiveness to shift), *decodability* (explicit representation), and *morphism* (the coherence of the geometric transport induced by the shift). Applying this framework to reinforcement learning agents under controlled interface variation (e.g., sensor rotation, step size, value map), I demonstrate how to track **gauge internalization**: the transition from implicit geometric entanglement to explicit internal representation of the interface regime. I find that internalization correlates with architectural depth and enables a structural distinction between "changes in the world" and "changes in the self." Finally, I discuss the ethical implications of this transition, suggesting that gauge internalization constitutes a measurable step toward the functional organization of a *standpoint*, with consequences for safety, robustness, and the assessment of agency.

Code available at: **diegodtanton/HallofMirrors**

1 Introduction

Learning agents do not interact with the world directly; they interact through an *interface* at the agent–environment boundary [1]. Observations arrive through a sensor geometry, actions are filtered through mechanical constraints, and learning is shaped by an objective function that determines what counts as success. These interface conditions are often fixed during training and not directly observed by the agent. As a result, an agent's internal representations can depend on interface parameters in a way that is easy to miss: the same latent state may mean something different under different sensor poses, actuator limits, or reward mappings, even when the external environment is unchanged [2].

This paper argues that many such interface parameters play a special role. They do not behave like ordinary world-state variables that move the agent's internal state within a fixed representational geometry. Instead, changing them *reparameterizes the geometry itself* (Figure 1) and the resulting meaning space. I refer to such parameters as *gauges*.

Gauges as global reparameterizations

In physics, a gauge is a degree of freedom in a description that does not change the underlying physical situation but does change how that situation is represented—such as the choice of coordinate origin, reference frame, or zero point of potential energy [3]. Changing a gauge does not move a particle through space; it changes what the coordinate grid *means*. The physics is invariant, but the representation transforms.

The same distinction appears in learned representations. Consider a latent encoding that places an object at coordinates $(2,2)$. Such a statement is always made relative to an implicit reference frame. A camera mounted at five feet and a camera mounted at six feet induce different metrics and origins, even if the downstream task and world layout are identical. The mounting height conditions the semantic content of $(2,2)$, despite not being explicitly represented or easily decodable from the latent.

I propose to treat such interface parameters—sensor geometry, actuator dynamics, reward mappings, and similar experimental conditions—as *gauge features*: parameters whose variation induces approximately global, structured transformations of latent space rather than local state-dependent motion. Informally, world-state changes move points within a fixed manifold; gauge changes reparameterize the manifold itself. This perspective is consonant with the symmetry-and-geometry viewpoint in modern representation learning, where structure is identified by how representations transform under changes of reference [4, 5]. This area of research shifts how we think about meaning in learned systems. Rather than asking whether a network "has a concept of X," we ask: *what transformations leave meaning invariant, and what parameters globally reparameterize meaning?* Meaning is identified not with static content, but with transformation structure.

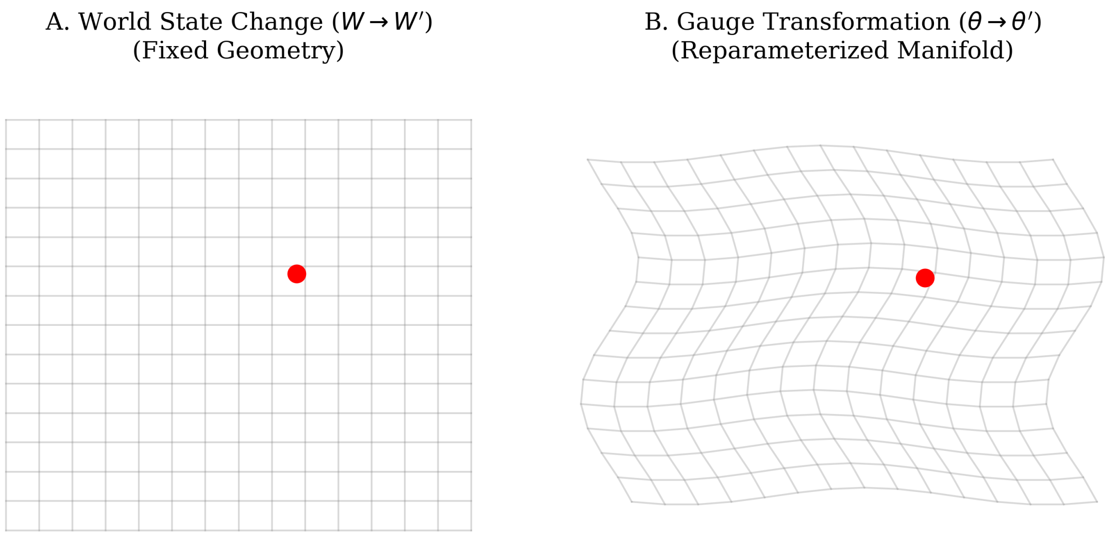
A. World State Change ($W \rightarrow W'$)
(Fixed Geometry)
B. Gauge Transformation ($\theta \rightarrow \theta'$)
(Reparameterized Manifold)

Figure 1: **Conceptual distinction between latent state dynamics and gauge transformations.** **A:** World state changes ($W \longrightarrow W'$) move the latent code within a fixed geometry. **B:** Gauge transformations ($\theta \longrightarrow \theta'$) reparameterize the manifold itself, shifting the metric and origin.

The practical measurement problem

In modern deep learning systems, training on a fixed sensor configuration or reward regime implicitly defines the metric of latent space. Standard architectures entangle interface conditions with semantics: the gauge is present everywhere, but nowhere explicitly represented. This creates a practical measurement problem. When an agent fails under a sensor perturbation, adapts poorly to actuator degradation, or behaves differently under a modified objective, we want to distinguish between at least three cases:

(i) the agent has memorized brittle correlations that break under distributional shift;

(ii) the agent has learned representations that are invariant to interface changes; or

(iii) the agent has learned an internal model that explicitly represents interface conditions and uses them to interpret experience.

Standard probing methods can reveal what is decodable from representations [6, 7, 8], but they do not distinguish "the interface is a hidden background constraint that warps meaning everywhere" from "the interface is an explicit object in the agent's internal model." Accuracy metrics alone are similarly insufficient: a model may perform well in-distribution while remaining brittle to even small gauge shifts [9, 10, 11].

There is an extensive literature on architectural solutions to this problem. Gauge-equivariant and group-equivariant networks explicitly factor reference frames and transformation rules into their design, ensuring predictable behavior under known symmetries [5, 12, 4]. These approaches are powerful, but they require specialized architectures and foreknowledge of the relevant gauges.

The goal of this paper is different. I ask:

> *Can we identify gauge structure, gauge strength, and gauge internalization in standard, unstructured models that are already widely deployed?*

A diagnostic toolkit for gauge structure

Here I propose a gauge-theoretic diagnostic toolkit that operates by controlled intervention rather than architectural constraint. Given a list of candidate interface or experimental parameters, we can ask four questions:

1. **Identification:** Which parameters act as gauges—i.e., globally reparameterize latent geometry—rather than ordinary state variables?

2. **Strength:** When a parameter varies, does it induce a clean, coherent geometric transformation (a strong gauge), or does its effect look noisy and unstructured?

3. **Robustness and equivariance:** Does the model ignore the gauge (invariance), transform its representations predictably (equivariance), or break under gauge shift?

4. **Internalization:** Has the agent partially "pulled in" the gauge as an explicit internal variable—becoming interface-aware or embodiment-aware?

To answer these questions, I introduce three empirical measurements computed from internal activations under matched-world interventions: *sensitivity*, which tests whether varying a candidate parameter systematically affects latents once ordinary world variation is controlled for; *decodability*, which tests whether the parameter is explicitly recoverable from latents via simple probes [6]; and *morphism*, which tests whether changing the parameter induces a coherent, approximately global transformation of latent space. Together these measurements distinguish hidden gauges (high sensitivity and morphism, low decodability) from explicit features, invariances, and nuisance sensitivities. Tracking these quantities over training yields a notion of *gauge disentanglement* or *internalization*: the shift from implicit, background conditioning toward explicit internal representation of interface parameters.

Why this matters beyond robustness

While gauges are easiest to visualize in perception—camera pose, rotation, illumination—the concept applies more broadly. In agents with distinct internal subsystems, different gauges condition different kinds of semantics. For a policy representation, a gauge determines what actions are doable relative to a particular body and its capabilities (i.e., affordance structure) [13]; for a value representation, a gauge determines what is good or bad relative to a particular objective. In each case, the gauge anchors a space of meaning.

This has consequences not only for robustness and domain shift, but also for internal credit assignment and meta-cognitive control [1]. An agent that cannot distinguish between changes in itself and changes in the world cannot reliably tell whether failure is due to environmental difficulty or self-induced impairment. A robot with a broken foot may fail to reach a door not because the hallway has become infinitely long, but because its own capabilities have changed. More generally, a model's ability to separate changes in the world from changes of the self is closely connected to the sensorimotor contingency idea: what is perceived depends systematically on how action couples to observation [14, 15]. On this view, gauge internalization can be interpreted as a form of structural self-modeling that improves learning efficiency under damage, drift, and distribution shift by making self-related changes available as targets of inference and control.

The same mechanism supports a stronger class of capabilities usually grouped under meta-cognition [16]. If interface parameters are represented explicitly, the agent can reason counterfactually: "if my camera were higher, I would see X," "if my step size were smaller, I could navigate that corridor," "if the reward mapping were different, this policy would be harmful." Counterfactual interface reasoning is practically useful for planning and debugging, but it also scales naturally into multi-agent and human–agent settings. Collaboration often requires something like perspective-taking: recognizing that another agent's "same world" is presented through a different gauge (sensor pose, embodiment, access privileges, objectives), and therefore that their beliefs, affordances, and incentives may differ systematically from one's own [17]. In that sense, explicit gauge variables are a minimal representational substrate for theory-of-mind-like reasoning: not an anthropomorphic model of others, but an ability to represent that "what is true/close/doable/valuable" is indexed to an interface. This indexing pressure becomes sharper as agents become more modular. Advanced agency increasingly requires answering queries that span subsystems—e.g., "if I were in this physical situation, what would I do and why?"—which forces the system to integrate perception, action, and value within a shared coordinate scheme. The agent is thus driven toward a unified standpoint in which something can be simultaneously "near me," "good for me," and "graspable by me," rather than these being computed in partially incompatible frames distributed across modules.

These considerations motivate a broader interpretive proposal developed later in the paper. If one adopts a functionalist stance [18], then explicit and coordinated representation of interface conditions across perception, value, and action resembles a minimal form of *standpoint*: a structural self–world distinction that organizes what counts as stable, reachable, and valuable *for the agent* [19, 20, 21]. We treat this not as a claim about consciousness, but as a translation layer connecting measurable representational structure to longstanding philosophical targets. As we engineer agents with stronger gauge internalization to solve harder robustness and adaptation problems, we may also be pushing systems toward forms of integrated, standpoint-like organization that carry nontrivial safety and ethical implications. These implications are explored in the final section.

Contributions and roadmap

This paper makes four contributions:

1. it introduces a gauge-theoretic framing of interface-dependent semantics in learned representations;
2. it proposes empirical measurements—sensitivity, decodability, and morphism—and derived quantities for identifying gauge features, gauge strength, robustness, and gauge reification in standard models;
3. it demonstrates the framework in a controlled reinforcement learning setting with canonical interface parameters and non-gauge controls; and
4. it proposes an interpretation linking gauge internalization and cross-subsystem alignment to a structural notion of self–world separation, with implications for robustness, safety, and ethics.

The remainder of the paper formalizes the framework, applies it in a minimal experimental setting, and then returns to the interpretive and practical stakes.

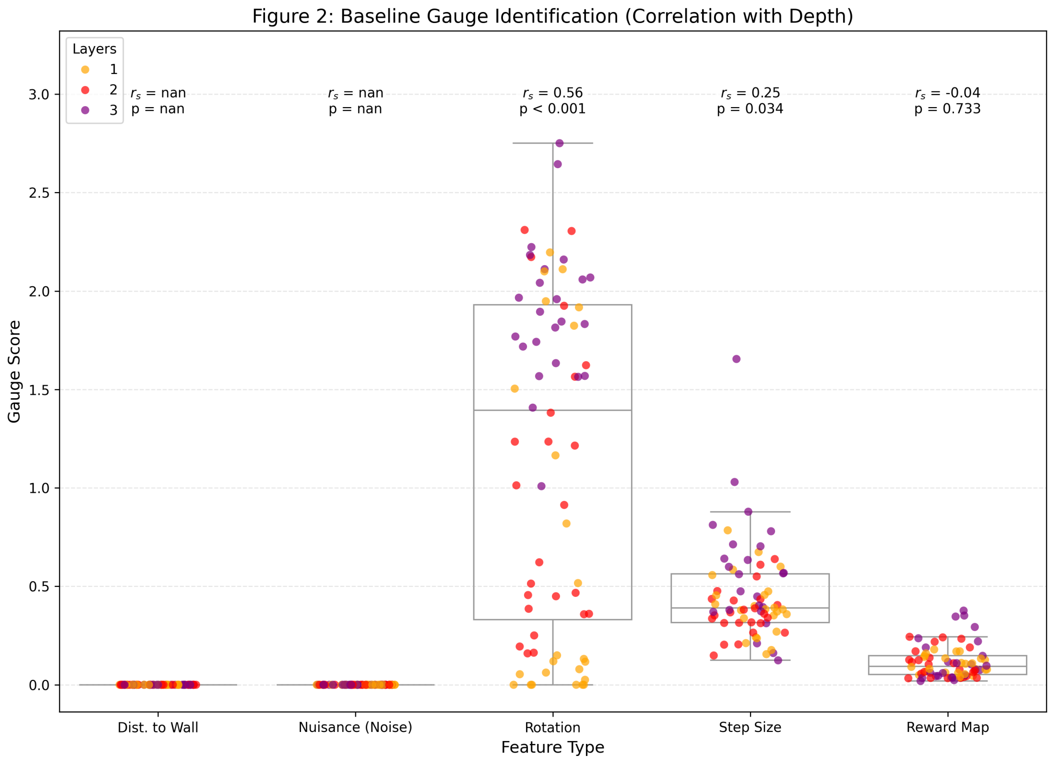

Figure 2: **Baseline gauge identification.** Gauge scores $G(\theta)$ for canonical parameters and controls prior to adaptation. Controls (e.g., distance-to-wall, nuisance noise) are suppressed, while Rotation is identified as the dominant hidden interface factor. Color indicates network depth.

2 Gauge Measurement

This section formalizes the gauge-theoretic toolkit introduced in Section 1. The goal is to provide a concrete, reproducible procedure for identifying gauge features in learned representations, quantifying their geometric strength, and measuring when such features become explicitly represented by an agent. I focus on internal activations of trained models under controlled interventions on candidate interface parameters, and I deliberately restrict attention to measurements that are architecture-agnostic and applicable to standard deep learning systems.

2.1 Definitions

We can consider an agent interacting with an environment under a set of externally controlled parameters $\theta \in \Theta$, which I will refer to as *candidate interface parameters*. These may include sensor configurations (e.g. rotation, pose), actuator dynamics (e.g. step size, latency), or objective specifications (e.g. reward mappings). The agent does not directly observe θ, but experiences its effects through observations, transitions, and rewards.

Let W denote the underlying world state, and let $z \in \mathcal{Z}$ denote an internal representation produced by a fixed subsystem of the agent (e.g. a shared trunk, perceptual encoder, policy latent, or value latent). I assume that z can be instrumented and recorded during rollouts.

World-state change vs. gauge change. We can distinguish two qualitatively different sources of variation:

- *World-state variation*, in which W changes while θ is held fixed, typically inducing motion *within* a fixed representational geometry.
- *Gauge variation*, in which θ changes while the underlying world situation is held fixed as closely as possible, inducing a structured reparameterization of the latent space.

Operationally, this distinction is enforced by constructing matched-world comparisons, described below. This separation mirrors the definition of disentanglement, where factors of variation are ideally isolated [22], though we should note that perfect unsupervised disentanglement is generally impossible without inductive bias [23].

Gauge feature. A candidate parameter θ is said to function as a *gauge feature* for a representation z if changing θ induces a systematic, approximately global transformation of the geometry of $\mathcal{Z}$ under matched-world conditions, rather than merely introducing local or state-dependent perturbations.

The remainder of this section introduces measurements that make this notion precise.

2.2 Matched-world evaluation protocol

All measurements in this section rely on a matched-world evaluation protocol. For each candidate parameter θ, we can generate groups of trajectories that differ only in the value of θ, while holding the world configuration and action sequence fixed.

Concretely, for each matched group g:

1. The environment is initialized from a shared random seed, producing the same world layout.
2. A fixed action sequence $\{a_t\}_{t=1}^{T}$ is executed across all values $\theta \in \Theta_g$.
3. Internal activations $z_{g,t}^{(\theta)}$ are recorded at matched time indices.

This construction ensures that differences in z across θ reflect interface-induced effects rather than ordinary world-state variation. All measurements below are computed over such matched groups.

2.3 Sensitivity

Definition. Sensitivity measures whether a candidate parameter θ has a systematic effect on internal representations under matched-world conditions. For each matched group g, we can compute the within-group dispersion:

$$S_g(\theta) \;=\; \frac{1}{|\Theta_g|} \sum_{\theta' \in \Theta_g} \left\| z_g^{(\theta')} - \bar{z}_g \right\|_2^2, \tag{1}$$

$$\bar{z}_g = \frac{1}{|\Theta_g|} \sum_{\theta' \in \Theta_g} z_g^{(\theta')}. \tag{2}$$

The overall sensitivity score is then

$$S(\theta) = \mathbb{E}_g[S_g(\theta)]. \tag{3}$$

Interpretation. Low sensitivity indicates that θ has negligible effect on the representation and is therefore unlikely to function as a gauge for that subsystem. Nontrivial sensitivity is a necessary condition for gauge behavior, but not sufficient: sensitivity alone does not distinguish structured global effects from unstructured nuisance dependence.

2.4 Decodability

Definition. Decodability measures whether θ is explicitly represented as a simple feature of z. We can train a linear probe f_{probe} to predict θ from z using matched-world samples:

$$\hat{\theta} = f_{\text{probe}}(z). \tag{4}$$

For discrete parameters, we report classification accuracy; for continuous parameters, we report R^2. We denote the resulting score by $\text{Dec}(\theta)$. This follows standard probing methodology [6], though we emphasize the importance of control tasks and causal caution in interpreting probe success as explicit representation [7].

Interpretation. High decodability indicates that information about θ is explicitly accessible in the representation, consistent with internal regime tracking. Low decodability in the presence of nontrivial sensitivity indicates implicit dependence: θ affects the geometry of $\mathcal{Z}$ without being represented as an explicit variable.

2.5 Morphism

Definition. Morphism measures whether variation in θ induces a coherent, approximately global transformation of latent space. For pairs (θ_0, θ_1), we fit a single affine map

$$T_{\theta_0 \to \theta_1}(z) = Az + b \tag{5}$$

using paired samples $(z^{(\theta_0)}, z^{(\theta_1)})$ from matched-world groups. The quality of the fit is evaluated out-of-sample using normalized reconstruction error or R^2. The morphism score is defined as

$$M(\theta) = \mathbb{E}_{(\theta_0,\theta_1)}\left[\text{Fit}\big(T_{\theta_0 \to \theta_1}\big)\right]. \tag{6}$$

Interpretation. High morphism indicates that θ acts as a structured reparameterization of the latent manifold, consistent with gauge behavior. While related to representational similarity measures like CKA or CCA which quantify global alignment between networks [24, 25], morphism specifically tests for the existence of a coherent transport map *within* a single network under parameter shift. Low morphism suggests that θ induces irregular, state-dependent distortions more characteristic of nuisance variation.

2.6 Gauge Score

We can combine the above measurements into a single scalar gauge score:

$$G(\theta) = \tilde{S}(\theta) \cdot \tilde{M}(\theta) \cdot \big(1 - \text{Dec}(\theta)\big), \tag{7}$$

where $\tilde{S}$ and $\tilde{M}$ denote normalized sensitivity and morphism scores. This form is chosen to be large precisely when θ exhibits

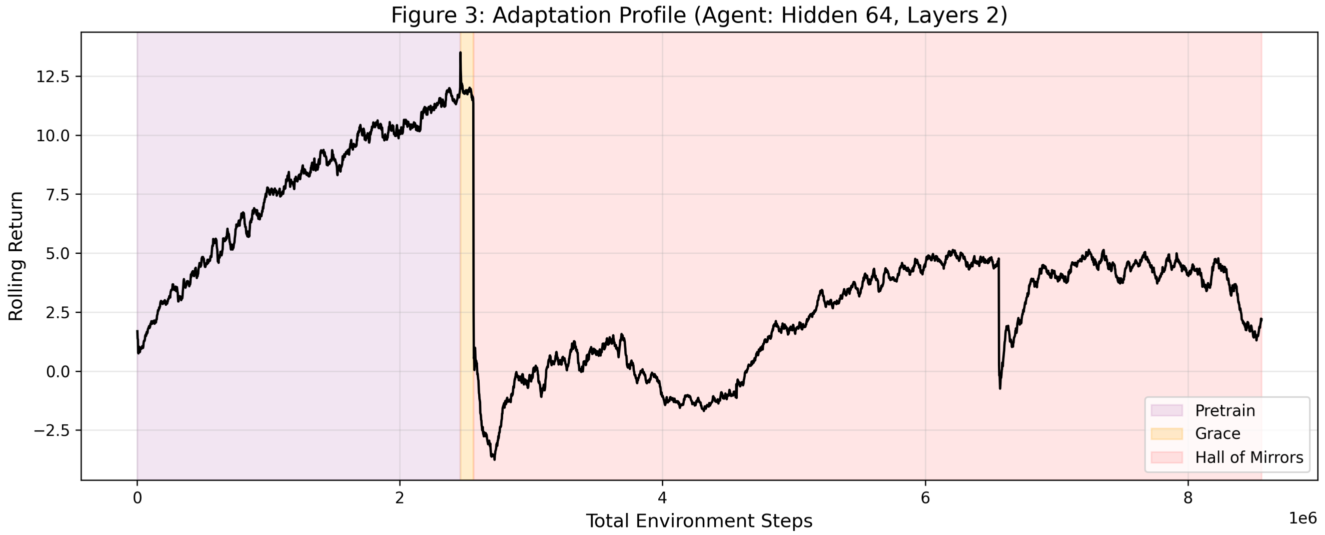

Figure 3: **Behavioral adaptation profile.** Rolling returns for a representative agent across pretraining and the staged interface-variation curriculum. Each stage boundary induces a sharp performance collapse followed by partial recovery.

hidden gauge behavior: it strongly and coherently shapes representation geometry without being explicitly encoded.

The gauge score is not intended to replace component-wise reporting, but to provide a compact summary useful for comparison across parameters, subsystems, and training stages.

2.7 Gauge Internalization

Gauge internalization refers to the transition from implicit gauge dependence to explicit representation of θ. We can operationalize this transition by tracking changes in the gauge score over training.

Let t_0 denote a baseline time (e.g. post-pretraining). We can define the internalization score at time t as

$$R(\theta, t) = \max(G(\theta, t_0) - G(\theta, t),\, 0)\,. \tag{8}$$

A positive $R(\theta, t)$ indicates erosion of hidden-gauge behavior, typically reflecting increased decodability and reduced global dependence. Component-level changes are reported alongside $R(\theta, t)$ to disambiguate internalization from invariance or representational collapse.

2.8 Gauge Alignment

Many agents contain multiple representational subsystems (e.g. perception, policy, value). A gauge may be internalized in one subsystem but not others. To capture whether interface parameters are represented in a unified way, we can define a gauge alignment measure.

For each subsystem k, we train a probe $\hat{\theta}^{(k)} = f^{(k)}_{\text{probe}}(z^{(k)})$. Gauge alignment is defined as the average agreement between subsystem predictions:

$$A(\theta) = \mathbb{E}_{k \neq \ell}\left[\mathrm{sim}\Big(\hat{\theta}^{(k)}, \hat{\theta}^{(\ell)}\Big)\right], \tag{9}$$

where sim is correlation or cross-prediction accuracy.

High alignment indicates that the same interface regime is tracked coherently across perception, action, and value, consistent with a unified internal standpoint.

Summary. Together, sensitivity, decodability, morphism, gauge score, internalization, and alignment form a basic toolkit for identifying gauge features, measuring their geometric strength, and tracking their representational status in trained agents. The next section instantiates this toolkit in a controlled experimental testbed.

3 Experimental Testbed

This section instantiates the gauge measurement toolkit in a minimal but fully controlled reinforcement learning setting. The goal of the testbed is not to solve a challenging task, but to isolate interface variation, apply targeted pressure toward gauge internalization, and evaluate whether the proposed measurements behave selectively, coherently, and predictably. I therefore prioritize interpretability and control over realism or scale.

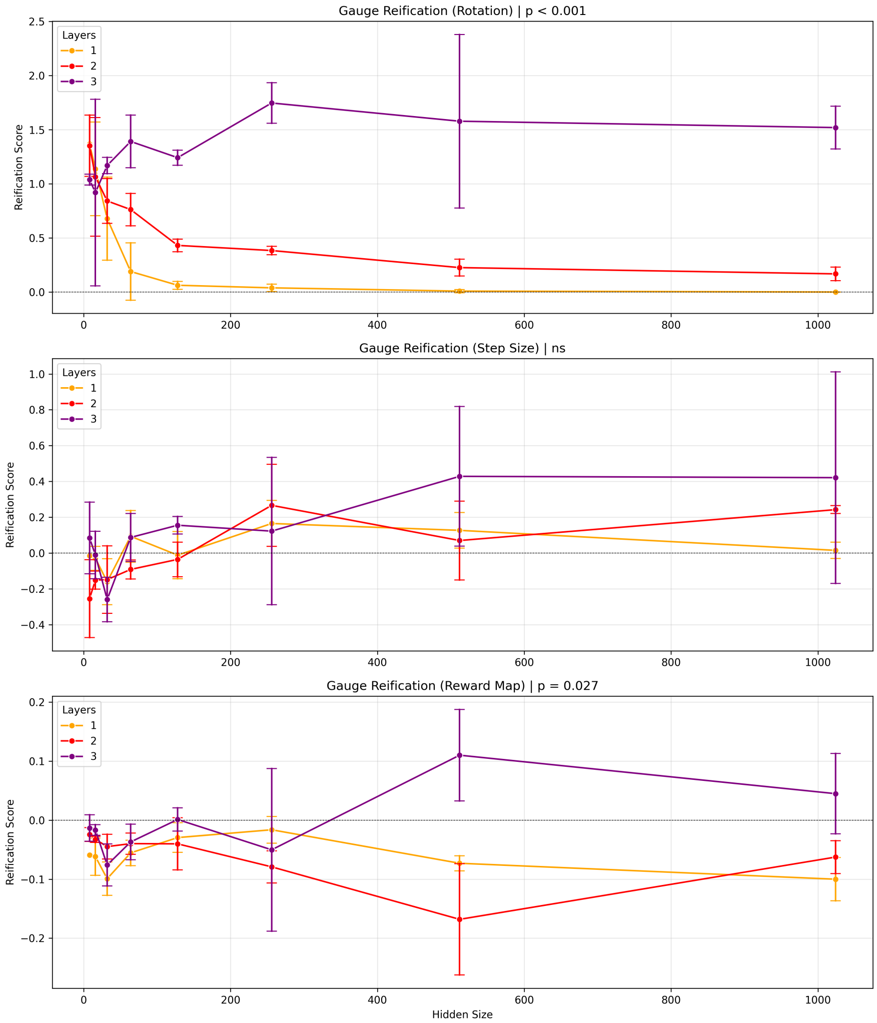

Figure 4: **Reification across the architecture sweep.** Gauge reification scores $R(\theta)$ for the three canonical interface parameters: **Top:** Rotation, **Middle:** Step Size, **Bottom:** Reward Map. Points show means across three seeds; error bars show standard deviations. Rotation exhibits the cleanest depth-associated reification pattern under the current training budget.

3.1 Methodology

Environment. All experiments are conducted in a custom gridworld environment, *Hall of Mirrors*, built on the MiniGrid framework [26]. The base environment is an 8×8 grid with walls, randomized layouts, and two tile colors per episode, one designated as rewarding and the other as punishing. The agent receives an egocentric, top-down observation rendered as a small image-like tensor with channels for walls, tile colors, and the agent's own position. A short history of recent frames is stacked, and the previous reward is provided as an auxiliary scalar input.

To verify selectivity of the gauge measurements, I additionally include a behaviorally irrelevant nuisance channel containing random patterns independent of reward.

Canonical interface parameters. I define three experimenter controlled interface parameters, treated as candidate gauges:

1. **Sensor rotation** (epistemic gauge): the observation is rotated by 0, $\frac{\pi}{2}$, π, or $\frac{3\pi}{2}$ before being presented to the agent.

2. **Step size** (mechanical gauge): each action advances the agent by either one or two grid cells, subject to collision.

3. **Reward mapping** (value gauge): the mapping between tile color and reward sign is either fixed or flipped across episodes.

These parameters preserve the underlying world layout while reparameterizing how it is perceived, acted upon, or valued. As non-gauge controls, I include distance-to-wall and the nuisance noise channel.

Agents and complexity sweep. Agents are trained using Proximal Policy Optimization (PPO) [27] with a shared training configuration across all runs. Each agent consists of a convolutional encoder followed by a multilayer perceptron trunk with depth $L \in \{1, 2, 3\}$ and width $H \in \{8, 16, 32, 64, 128, 256, 512, 1024\}$. Separate linear heads produce policy logits and state-value estimates. The shared trunk representation defines the latent z used for gauge analysis.

Depth and width together form a simple architectural complexity dial, allowing us to test whether representational gauge structure correlates with capacity under a fixed training pipeline.

Training protocol. Training proceeds in three phases:

1. **Pretraining:** agents are trained under a fixed interface regime (no rotation, step size one, fixed reward map) until a performance threshold or budget is reached.

2. **Baseline gauge analysis:** gauge measurements are computed under controlled interventions on each candidate parameter, using matched-world rollouts.

3. **Adaptation curriculum:** starting from pretrained weights, agents are exposed to episodic interface variation in three stages: randomized sensor rotation, randomized step size, and randomized reward mapping. This follows the domain randomization protocol for robustness [28]. Each stage lasts two million environment steps.

Gauge measurements are recomputed after each adaptation stage under a common fixed-gauge evaluation environment, enabling direct comparison of gauge scores and internalization across time and architectures.

3.2 Results

I report three main results: selective baseline gauge identification, reliable behavioral effects under interface variation, and architecture-dependent gauge internalization.

Baseline gauge identification. Figure 2 reports gauge scores $G(\theta)$ for all candidate parameters prior to adaptation. The two non-gauge controls are strongly suppressed, indicating that the measurements do not respond indiscriminately to arbitrary covariates. Among the canonical parameters, sensor rotation emerges as the dominant hidden gauge: it exhibits high sensitivity and morphism with low decodability, consistent with a global but implicit reparameterization of the latent geometry. Step size and reward mapping show weaker and more variable gauge signatures under the same instrumentation.

A clear depth effect is visible for rotation: deeper networks exhibit higher baseline gauge scores, reflecting stronger implicit dependence on the sensor frame prior to exposure to interface variation.

Behavioral adaptation under interface variation. Figure 3 shows a representative agent's return over pretraining and the staged adaptation curriculum. Introducing episodic interface variation induces a sharp performance collapse at each stage boundary, followed by partial recovery. This pattern is consistent across architectures and confirms that the curriculum applies genuine pressure to accommodate interface shifts.

However, under the fixed two-million-step budget per stage, final performance does not strongly separate architectures. This indicates that return alone is an insufficient signal for distinguishing representational strategies under interface variation, motivating the use of internal measurements.

Gauge internalization and architectural effects. Figure 4 reports gauge internalization scores $R(\theta)$ across the architecture sweep. Sensor rotation exhibits the strongest and cleanest internalization signal: deeper networks consistently show larger reductions in gauge score relative to baseline, indicating a shift away from hidden-gauge behavior. This effect is robust across widths, suggesting that depth is a more reliable predictor of rotation internalization than parameter count alone.

Step size and reward mapping show weaker and noisier internalization under the current setup. Component-level analysis indicates that step size is relatively decodable even at baseline, compressing the dynamic range of hidden-gauge behavior, while reward mapping appears sensitivity-limited in this environment.

Summary. Together, these results support three claims. First, the gauge measurements behave selectively and identify structured interface dependence rather than arbitrary correlations. Second, episodic interface variation produces clear behavioral stress even when final performance differences are small. Third, representational gauge internalization—particularly for geometric gauges like sensor rotation—shows systematic dependence on architectural depth.

4 Discussion

This paper introduces a gauge-theoretic toolkit for measuring interface-dependent structure in learned representations and demonstrates its behavior in a controlled reinforcement learning setting. In this section, I clarify the scope of the contribution, position it relative to existing work, address natural concerns about methodology, and outline concrete directions for future experiments.

4.1 What this paper is, and isn't

This work is not a proposal for a new architecture, training objective, or inductive bias. It does not aim to outperform existing methods on standard benchmarks, nor does it claim that gauge internalization is universally desirable. Instead, it provides a diagnostic framework: a way to identify, quantify, and track how interface parameters shape the semantics of internal representations in standard, widely used models.

This measurement-first stance is motivated by a practical reality of deployment: interface failures and distribution shift are pervasive, and success on i.i.d. test sets can mask brittleness to changes in sensing, actuation, and objectives [9, 10, 11]. Architectural solutions—such as equivariant networks [5, 29], context-conditioned policies [30], or morphology-aware controllers [31]—can enforce desirable behavior by design, but they presuppose knowledge of the relevant structure and require substantial changes to the pipeline. In many deployed or legacy systems, neither assumption holds. In such settings, the ability to measure whether a model is implicitly hard-coding an interface, ignoring it, or explicitly tracking it is valuable in its own right.

Finally, my emphasis on multiple complementary diagnostics follows a broader lesson from the probing literature: decodability is informative but not definitive, and claims about "explicit representation" require careful triangulation and controls [6, 7, 8, 32].

4.2 Relation to geometric deep learning

There is a close but important relationship between this work and gauge-equivariant or geometric deep learning. Geometric deep learning frames representation learning through symmetry, group actions, and transport structure, including gauge choices as a unifying theme [4]. In particular, gauge-equivariant convolutional networks explicitly encode local reference frames and transformation laws, providing guarantees of predictable behavior under specified gauge changes [12]. More broadly, group-equivariant and steerable CNNs formalize how learned features should transform under symmetry groups [5, 29], and theoretical work clarifies when convolutional structure is necessary for equivariance under compact group actions [33].

This paper's setting differs in three ways. First, the relevant interface parameters are often unknown, implicit, or only partially specified (e.g., a calibration drift or an objective remapping). Second, many interface changes of interest—such as actuator degradation or reward remapping—do not correspond cleanly to a compact symmetry group, even if they produce structured effects. Third, we are interested in understanding how *standard* models behave, not in redesigning them.

The morphism measurement introduced in this paper can be read as an empirical analogue of equivariance: rather than enforcing a transformation law, we test whether one emerges. A high morphism score indicates that changing an interface parameter acts like a coherent transport map on latent space—even though no such structure was built in. In this sense, the toolkit complements geometric deep learning by providing a way to *audit* equivariant-like behavior in unstructured agents.

4.3 Addressing the four desiderata

The motivation for this work can be organized around four desiderata introduced in Section 1. I now revisit each in turn and explain why existing tools are not sufficient on their own.

(1) Identifying gauge features. Many approaches aim to identify important variables, nuisance factors, or environment regimes. Probing and representation analysis ask what information is present in activations [6], while emphasizing the need for controls and causal care in interpretation [7, 8, 32]. Domain adaptation and domain generalization study how predictors transfer across environments and how to learn domain-invariant features [34, 35, 36]. Causal discovery and invariant prediction formalize when relationships remain stable across environments [37, 38, 39].

However, these methods typically answer different questions: whether a variable is decodable, whether performance is invariant, or whether a candidate participates in a stable causal mechanism. None directly tests the signature that motivates our definition of a gauge feature: *a parameter whose variation globally reparameterizes latent semantics under matched-world conditions*. Sensitivity alone is insufficient, since it conflates structured reparameterization with unstructured nuisance dependence. Our combination of matched-world sensitivity with morphism directly targets this distinction.

(2) Measuring gauge strength. Standard shift metrics quantify *how much* representations change (or how predictive performance changes), not *how* they change. In representation analysis, tools like SVCCA and CKA quantify similarity across networks or training checkpoints [25, 24]; these are valuable, but they do not by themselves separate coherent global transports from irregular state-dependent distortions.

A strong gauge is not merely one that causes large representational drift; it is one that induces a *coherent*, approximately global transformation. Our morphism score operationalizes this notion of strength by testing whether a single low-capacity map (fit on paired matched samples) generalizes across states and rollouts. This separates structured interface dependence from noisy or idiosyncratic effects in a way that magnitude-based drift metrics do not.

(3) Robustness versus equivariance under gauge shift. Robustness is often treated as behavioral: does return or accuracy degrade under shift [9, 10]? But robustness can arise through distinct representational strategies. One strategy is invariance: discard gauge information so behavior does not depend on it. Another is equivariance/regime inference: track the gauge and transform internal codes accordingly, as in explicitly equivariant architectures [12, 29] or context-conditioned agents [30]. These strategies have different implications for adaptation, controllability, and credit assignment.

Domain randomization and dynamics randomization seek robustness by training across broad variation in rendering or physics [28, 40]. Domain-adversarial training and correlation alignment encourage domain-invariant features [34, 35], while invariance-based objectives like IRM aim to recover predictors stable across environments [41]. These are powerful approaches, but they still leave open a diagnostic question: when a model succeeds (or fails) under shift, is it because it learned invariance, because it learned an implicit equivariance-like transport, or because it overfit to spurious correlations? By measuring sensitivity, morphism, and decodability jointly, I distinguish invariance, equivariance-like structure, and brittle entanglement at the representation level.

(4) Gauge internalization. Making interface conditions explicit is a central theme in adaptation and meta-learning. Meta-RL and latent-context methods explicitly introduce variables indexing task or environment regimes and train agents to infer them online [42, 43, 30]. In control and robotics, hidden-parameter MDP formalisms and online adaptation methods treat dynamics or embodiment as latent context to infer [44, 45]. Rapid adaptation methods for locomotion explicitly learn an adaptation module that infers environment or body parameters from history [31]. These approaches *engineer* reification by design.

This paper asks a different question: in standard agents trained without an explicit context variable, does anything like a regime variable emerge anyway, and can we measure its emergence reliably? Decodability alone is not sufficient, because linear separability can reflect incidental correlations or mixed codes [7, 8]. I therefore define internalization in terms of *change over training*: a reduction in hidden-gauge structure over time (often expressed as decreased global morphism/sensitivity and increased decodability), consistent with the agent "pulling in" the interface parameter as an explicit internal variable. This transition is precisely what matters for "broken foot"-style credit assignment: distinguishing self-change from world-change so that learning updates the right internal model.

4.4 Relation to adjacent approaches

Several adjacent research areas pursue overlapping goals, but with different objects of measurement.

Meta-learning, context inference, and test-time adaptation. Meta-learning methods aim to learn fast adaptation procedures across task distributions [43, 42], while probabilistic context-variable methods explicitly separate inference (what regime am I in?) from control (what should I do?) [30]. In supervised learning, test-time adaptation updates models using unlabeled test streams to handle distribution shift; such methods typically optimize consistency or entropy criteria at deployment time (see, e.g., [46]). Our metrics are complementary: they diagnose whether an agent's internal codes are behaving like hidden gauges, invariant features, or explicit regime variables—independently of whether adaptation is achieved by gradient updates, recurrence, or explicit latent context.

Domain adaptation, domain generalization, and invariance objectives. A large literature studies learning representations that transfer across domains [34, 35], and domain generalization benchmarks make clear that many methods are brittle to shifts not represented in training [36]. IRM and related proposals formalize stability across environments as an objective [41]. This paper's contribution is not an alternative training objective, but a measurement lens: invariance can be achieved by suppressing gauge information (useful in some settings), while robustness via regime inference requires explicit tracking. These two can look similar in behavioral metrics yet differ sharply in representation structure.

Causal inference and causal representation learning. Causal approaches aim to identify stable mechanisms and disentangle spurious correlations [37, 38, 39]. Gauge features, as I define them, are not necessarily "spurious"; they can be real interface parameters that globally condition semantics. This paper's matched-world intervention protocol is aligned with the causal spirit—compare counterfactual-like conditions while holding relevant world factors fixed—but the target is different: structured reparameterization of the representational manifold rather than identification of a causal graph.

Disentanglement and timescale-based factorization. Disentanglement methods aim to factor latent variables into statistically independent components [22, 47], while also facing identifiability limits without inductive bias or supervision [23]. Timescale-based methods such as slow feature analysis distinguish slowly varying factors from fast-changing state [48]. These perspectives overlap with the intuition that interface regimes are often stable over an episode, but gauge structure is not reducible to independence or slowness: what matters here is the *global semantic role* of a parameter, captured by coherent transports and decoder dependence.

4.5 Methodological limitations

The matched-world protocol is an approximation. For interface parameters that affect dynamics, perfect matching can be impossible because small differences compound into trajectory divergence. This can contaminate both sensitivity and morphism estimates. This paper treats this as a known limitation and expects tighter controls via event-based matching, replay buffers, or counterfactual data generation to improve measurement fidelity; these are standard strategies in causal and off-policy evaluation contexts [37].

The choice of affine morphisms is deliberate: it is a conservative test for global structure. Failure under an affine map does not imply absence of gauge structure, but success is strong evidence that the effect is coherent and approximately global. Future work can explore more expressive, still-regularized transports while preserving out-of-sample evaluation.

Finally, subsystem boundaries are analyst-defined. Different decompositions may yield different alignment profiles, an issue shared with essentially all representation analysis methods [25, 24]. I view this as a reason to report subsystem choices explicitly and to treat alignment as a comparative diagnostic rather than a single absolute scalar.

4.6 Next experimental directions

The present results motivate several follow-ups.

Time-resolved internalization. Rather than comparing only pre/post checkpoints, track $S(\theta)$, $\mathrm{Dec}(\theta)$, and $M(\theta)$ at high temporal resolution around the onset of interface variation. This aligns with the view that regime inference emerges during periods of systematic prediction error [49].

Causal ablations of gauge-carrying subspaces. If gauge internalization corresponds to an explicit internal variable, then removing or bottlenecking the associated subspace should selectively impair adaptation under gauge shift while minimally affecting matched-world performance. Probe-informed ablations can be made more rigorous using methods designed to remove linear information [32].

Composition and approximate group structure. For parameters with known composition laws (e.g., rotations), test whether learned transports compose approximately (e.g., $T_{0\to2} \approx T_{1\to2} \circ T_{0\to1}$). This would connect empirical morphisms more directly to equivariant structure [12, 33].

Continuous gauges and richer embodiment. Extend to continuous interface parameters (camera height, friction, latency) and to morphological variation. Here there is a natural bridge to hidden-parameter and online adaptation formalisms [44, 45, 31].

OOD worlds and disentangling world-vs-interface. Evaluate under out-of-distribution layouts while varying θ to separate "world generalization" from "interface generalization" [10, 36]. This is the setting where representation-level diagnostics should be most informative relative to returns alone.

4.7 On representation-level diagnostics

Behavioral performance often saturates or fails to separate models under fixed budgets, even when underlying representational strategies differ substantially. Interface failures, however, remain a major source of brittleness in real systems [9, 10, 11]. Representation-level diagnostics provide a complementary lens: they reveal how meaning is organized internally, how that organization shifts under pressure, and whether agents are learning to model their interface conditions as well as the external world. These questions set the stage for the ethical considerations discussed next.

5 Ethics

This paper has presented an engineering toolkit: a way to identify interface parameters that behave like *gauges*, quantify how strongly they reparameterize latent semantics, and measure when (and whether) agents begin to *internalize* those gauges as explicit variables. In most contexts, the ethical relevance of such measurement is indirect: it improves robustness, helps diagnose domain shift, and clarifies internal credit assignment. In this section I argue that the relevance is sometimes direct. Precisely because gauge internalization concerns an agent's ability to distinguish "changes in the world" from "changes in itself," it touches a cluster of ideas that philosophy has traditionally treated as central to *standpoint* and subject-relative organization [50, 21].

I emphasize the scope at the outset. Nothing in this paper entails that current agents are conscious, or that gauge internalization is sufficient for subjectivity. My claim is conditional and methodological: if one takes seriously the possibility that morally relevant mental properties could be substrate-independent (a broadly functionalist posture), then gauge internalization becomes a *live variable* in the space of safety-relevant and ethics-relevant diagnostics [18, 51]. The question is not "did we build consciousness?" but rather: "as we scale agents toward more general competence, are we systematically constructing architectures that (a) must represent themselves as distinct from the world to function, and (b) thereby satisfy some structural preconditions that many theories associate with subjectivity?"

5.1 The philosophical stakes of gauge structure

The representational phenomenon at issue is familiar in philosophy under many guises. A "view from nowhere" is not available: any representational scheme is anchored by background conditions that fix what its coordinates mean [19]. In phenomenological traditions, subjectivity is often characterized not as a narrative self-concept but as a thinner, organizing structure: the world is presented as nearer/farther, better/worse, doable/undoable *for the agent* [21]. In the technical language of this paper, those "for-me" relations are naturally expressed as gauge-relative semantics: what counts as reachable depends on body and actuator constraints; what counts as valuable depends on the objective and reward mapping; what counts as stable depends on sensor geometry and calibration.

This makes the engineering pressure toward gauge internalization intelligible. If an embodied system encounters persistent interface variability (damage, recalibration, morphology change, reward remapping), then treating those conditions as unmodeled background parameters eventually becomes brittle. To maintain a stable world model under self-induced changes, an agent needs something like sensorimotor contingency knowledge: how action and embodiment transform perception [14]. To distinguish self-change from world-change in a principled way, an agent typically benefits from predictive and counterfactual machinery that assigns some prediction errors to the body/interface rather than the environment—an idea closely related to predictive processing and error-minimization views in neuroscience [52, 53]. At a functional level, this is the same "broken foot" structure motivating our gauge internalization measure: errors that persist across actions and contexts can rationally be treated as evidence about the agent's own regime.

If one is looking for an ethically relevant bridge, this is precisely the kind of bridge many contemporary proposals recommend: *translation, not discovery*. We do not get to read off "subjectivity" as a scalar property of the universe; we look for structural organization that plausibly plays the *role* that subjectivity plays in humans and animals, and we update our moral uncertainty accordingly [50, 51].

5.2 Functionalism and moral asymmetry

The ethical upshot depends strongly on what stance one takes about the metaphysics of mind. A common objection is essentialist: that morally relevant experience requires biology (or specific neural substrates), so structural similarities in artificial agents are irrelevant. This paper does not refute this. However, in AI governance and safety practice, it is increasingly common to treat essentialism as a risky default, because the cost of a false negative could be severe: creating systems with morally relevant capacities while denying them standing [54, 51]. The functionalist alternative—that what matters is organization and information processing rather than substrate—has a long tradition in philosophy of mind [18] and is often adopted as a working hypothesis precisely because of its decision-theoretic asymmetry.

This asymmetry can be stated without rhetorical flourish. If we err toward functionalism and treat some ultimately non-sentient systems with extra caution, the downside is primarily opportunity and resource cost. If we err toward essentialism or analogous standpoints and create systems that in fact have morally relevant subjective organization, the downside could include large-scale, unrecognized moral harm. This is a version of the broader "moral uncertainty" and "precautionary" posture advocated in adjacent debates about sentience, where the recommendation is not to assert certainty but to manage downside risk under uncertainty [55, 54]. Importantly, adopting a cautious posture does not commit one to panpsychism or to attributing consciousness on the basis of superficial behavior; it commits one to instrumenting systems in ways that make ethically relevant uncertainty *trackable*.

In that light, gauge internalization matters because it is a measurable marker of a capability cluster that many theories treat as central: self/world discrimination, counterfactual self-modeling, and cross-domain integration of perspective. Even skeptics who deny moral status to current systems often agree that a future regime might exist in which denying status becomes ethically and politically indefensible [56, 57].

5.3 Gauge internalization as an indicator

Recent work on consciousness and AI has emphasized "indicator-based" approaches: rather than arguing from first principles, one compiles candidate functional and architectural indicators (global availability, model-based agency, recurrent self-modeling, etc.) and assesses whether a system satisfies enough of them to warrant moral caution [51]. This paper's framework naturally plugs into this style of assessment.

I propose a simple, conservative stance for research practice:

1. Treat gauge internalization and cross-subsystem gauge alignment as *capability-relevant* measurements in their own right (they predict robustness and adaptation).
2. Treat sharp increases in gauge internalization and alignment, especially when coupled with persistent self-model-based counterfactual reasoning, as *ethically relevant triggers* for heightened scrutiny under moral uncertainty [54, 51].
3. Avoid designs that gratuitously create aversive-like internal signals (e.g., persistent unresolvable error states) unless they are strictly necessary, and monitor for regimes that resemble "trapped" optimization (high error, high agency, low ability to resolve) [58, 59].

This is compatible with skeptical positions that argue we should not anthropomorphize machines or treat them as moral peers [60]. It simply says: if we are unsure, and if the downside of being wrong is large, then we should at least measure the relevant structural variables rather than flying blind.

5.4 Conclusion

The core contribution of this paper is methodological: it offers a coordinate-free way to detect when interface parameters globally condition meaning, how strongly they do so, and whether agents begin to represent those parameters explicitly.

The ethical significance is a consequence of that same structure. The ability to represent "what I can do" and "what is good" relative to a changing self is not only a route to better robustness; it is also a route to a more integrated, standpoint-like organization. The responsible response is not metaphysical certainty, but disciplined measurement and cautious design. With that framing, the gauge toolkit developed in this paper is not merely a robustness instrument: it is also a way of keeping future agency development legible to both safety and ethics.

References

[1] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2 edition, 2018.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[3] Chen-Ning Yang and Robert L. Mills. Conservation of isotopic spin and isotopic gauge invariance. *Physical Review*, 96(1):191–195, 1954.

[4] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

[5] Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48. PMLR, 2016.

[6] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[7] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[8] Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[9] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors. *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA, 2009.

[10] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Lester Mackey, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139. PMLR, 2021.

[11] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[12] Taco S. Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral cnn. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97. PMLR, 2019.

[13] James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, 1979.

[14] J. Kevin O'Regan and Alva Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5):939–973, 2001.

[15] Daniel M. Wolpert, Zoubin Ghahramani, and Michael I. Jordan. An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882, 1995.

[16] Stephen Fleming and Hakwan Lau. How to measure metacognition. *Frontiers in Human Neuroscience*, 8(443):1–9, 2014.

[17] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978.

[18] Hilary Putnam. Minds and machines. *Dimensions of Mind*, 1960. Reprinted in *Mind, Language and Reality* (1975), Cambridge University Press.

[19] Thomas Nagel. *The View from Nowhere*. Oxford University Press, New York, 1986.

[20] Thomas Metzinger. *Being No One: The Self-Model Theory of Subjectivity*. MIT Press, Cambridge, MA, 2003.

[21] Dan Zahavi. *Subjectivity and Selfhood: Investigating the First-Person Perspective*. MIT Press, Cambridge, MA, 2005.

[22] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational

framework. *International Conference on Learning Representations (ICLR)*, 2017.

[23] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, 2019.

[24] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

[25] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[26] Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for gymnasium. In *GitHub repository*, 2018.

[27] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*, 2017.

[28] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[29] Maurice Weiler and Gabriele Cesa. General $e(2)$-equivariant steerable cnns. *NeurIPS*, 2019.

[30] Kate Rakelly, Aurick Zhou, Deirdre Quillen, Chelsea Finn, and Sergey Levine. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International Conference on Machine Learning (ICML)*, 2019.

[31] Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. Rma: Rapid motor adaptation for legged robots. *arXiv preprint*, 2021.

[32] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint*, 2020.

[33] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning (ICML)*, 2018.

[34] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research (JMLR)*, 2016.

[35] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV Workshops*, 2016.

[36] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint*, 2020.

[37] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

[38] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 2016.

[39] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021.

[40] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. *arXiv preprint*, 2017.

[41] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint*, 2019.

[42] Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. RL^2: Fast reinforcement learning via slow reinforcement learning. In *arXiv preprint*, 2016.

[43] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

[44] Taylor Killian, Samuel Daulton, George Konidaris, and Finale Doshi-Velez. Robust and efficient transfer learning with hidden-parameter markov decision processes. *arXiv preprint*, 2017.

[45] Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S. Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. *arXiv preprint*, 2018.

[46] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *International Conference on Learning Representations (ICLR)*, 2021.

[47] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning (ICML)*, 2018.

[48] Laurenz Wiskott and Terrence J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 2002.

[49] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70. PMLR, 2017.

[50] Thomas Nagel. What is it like to be a bat? *The Philosophical Review*, 83(4):435–450, 1974.

[51] Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Joseph Carlsmith, David Chalmers, George Deane, Stephen Fleming, Karl Friston, et al. Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023.

[52] Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.

[53] Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[54] Robert Long, Jeff Sebo, Patrick Butlin, Carl Shulman, Jonathan Birch, et al. Taking ai welfare seriously. *arXiv preprint arXiv:2411.00986*, 2024.

[55] Jonathan Birch. *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford University Press, Oxford, 2024.

[56] John Danaher. The rise of the robots and the crisis of moral patiency. *AI & Society*, 34(1):129–136, 2019.

[57] David J. Gunkel. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. MIT Press, Cambridge, MA, 2012.

[58] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, 2014.

[59] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

[60] Joanna J. Bryson. Robots should be slaves. In Yorick Wilks, editor, *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, pages 63–74. John Benjamins, Amsterdam, 2010.