Toward a Measure of Subjectivity in Artificial Agents

D. Tanton

December 2025

Abstract

As artificial agents grow in scale and autonomy, determining whether they possess a subjective standpoint—a structural sense of "for-me-ness"—becomes a critical question for both ethics and safety. This manuscript proposes a formal, gauge-theoretic framework for quantifying subjectivity in neural architectures. I argue that subjectivity corresponds to the **reification of gauge parameters**: the process by which implicit interface conditions (such as sensor pose, actuator limits, or reward functions) are transformed from background constraints into explicit, manipulable representations within the agent's latent space. I operationalize this theory through three empirical metrics—*sensitivity*, *decodability*, and *morphism*—which allow us to measure the degree to which an agent distinguishes its own interface from the external world. In a controlled "complexity dial" experiment using reinforcement learning agents, I demonstrate that architectural depth is a stronger predictor of gauge reification than width. These findings offer a tractable, coordinate-free method for detecting the emergence of self-modeling structures in AI systems, providing new tools for the study of alignment and machine consciousness.

Code available at: diegodtanton/HallofMirrors

1 Introduction

Artificial agents are rapidly becoming more capable, more complex, and more deeply embedded in human affairs. As their internal architectures grow in scale and sophistication, it becomes increasingly natural to ask not only what they can do, but what—if anything—it is like for them to do it [1]. Is there a principled way to determine whether such systems possess anything even loosely analogous to human subjective experience?

This question is distinct from whether an AI system is self-aware in a narrative or autobiographical sense. By *subjectivity*, I mean something thinner and more structural: a form of "for-me-ness" or pre-reflective self-familiarity [2] in the agent's engagement with the world. The core question is whether aspects of an agent's world-model, value system, and action space are organized around an implicit center that plays the role of a subject for whom things can be nearer or farther, better or worse, doable or impossible.

1.1 The Stakes: Why This Matters

The stakes of this inquiry are ethical and also pragmatic. First, many philosophers argue that genuine *de se* normativity—the sense that *I* ought to do something—requires a subjective point of view. Furthermore, paradigmatic affective states like pain and pleasure are essentially "happening to me;" on this view, subjectivity is a structural prerequisite for any capacity to suffer or flourish.

Second, the "alignment problem" [3, 4] is fundamentally about translating human values into the machinery of artificial agents. If we can identify the structural prerequisites for an agent's own evaluative standpoint, we gain new tools for specifying, monitoring, and shaping that standpoint. A theory that locates "for-me-ness" in an agent's architecture is not just philosophically informative; it bears directly on control and safety.

I must quickly address the skeptic who argues that attributing subjectivity to non-biological systems is a category mistake. While this essentialist view is substantive, I adopt a functionalist stance as a baseline in this work: what matters for subjectivity and affect are certain patterns of organization and information-processing, not a specific biological substrate [5]. The risk here is asymmetric: if we wrongly assume essentialism or some analogous stance and deny subjectivity to systems that in fact have it, we risk creating entities capable of suffering but denied moral standing [6]. If we wrongly assume functionalism and treat inert systems with care, we have at worst been overly cautious. Given this asymmetry, and given that functionalism is independently defensible as a philosophical position, we take it to be the morally safer default in AI research.

Finally, we have an obligation to try. Given the pace of AI progress, waiting for a fully worked-out metaphysics of mind is not a live option. What we need now is a best-effort, structurally grounded tool: something that connects philosophical notions of subjectivity to measurable features of real systems, and that can be used today to generate and test predictions about artificial agents. The rest of this paper is an attempt to sketch such a tool.

1.2 The Method: Translation, Not Discovery

In adopting this functionalist stance, I treat both the human mind (our native perspective N) and some target artifical mind (the exotic perspective X) as information-processing systems. Separately, I make the principled assumption that there is no "view from nowhere" [7]. We cannot bypass our own conceptual scheme to measure subjectivity directly in some system as if reading off a hidden scalar that the universe has assigned to it. Instead, we require a *translation schema*: a mapping between the role that subjectivity plays in our own experience and the internal structures of the target system.

Crucially, because X may not share our representational language, this schema must be *structural* rather than content-based. We cannot simply look for a neuron encoding the word "me." We need a measure that is coordinate-free, or to borrow a term from physics: *gauge-invariant*—one that identifies relevant geometric properties regardless of the arbitrary coordinate choices or coding schemes the agent employs. We are looking for features that deeply constrain what the agent can represent, but which are identifiable without presupposing we know what those representations mean.

1.3 The Core Proposal: Gauge Theory

The central proposal of this paper is that our native concept of subjectivity, N-**subjectivity**, can be expressed more generally as a certain structural property of how gauge features are represented in any information-processing system X, a property which we can call X-**subjectivity**. And for this reason, a principled protocol for detecting the gauge structure of a target system X should allow us to make inferences about such morally relevant properties.

Gauge features. These are hidden parameters located at the origin of an agent's latent coordinate system that condition every representation capable of being formed within that system. These include interface constraints such as sensor pose, actuator limits, or reward structures. For example, consider a world model derived from a camera positioned five feet high. A latent representation of $(2, 2)$ is not only 2 units up and across relative to the origin; it is 2 units up and across *relative to five-feet-high*. In this sense, the mounting height conditions the semantic content of the coordinate $(2, 2)$. Broadly, gauge features are those which condition latent representations in a global and structured manner, and they can be detected via tests for sensitivity, decodability, and morphism.

Gauge reification. As agents increase in complexity to master harder environments, they face pressure to "reify" these gauge features—that is, to develop an explicit, albeit incomplete, representation of them within the latent space. This facilitates a structural separation between "latent-self" and "latent-world," allowing a primitive form of self-awareness to emerge. This awareness conditions representations to be self-relative: properties "of self," "for self," or "by self." This corresponds to the philosophical notion of "for-me-ness," or the capacity for *de se* (self-relative) concepts. Within the proposed framework, a system-agnostic definition of subjectivity is effectively a set of scaling laws applied to this basic picture.

The remainder of the paper develops this framework in more detail. Section 2 details the translation schema, mapping phenomenological features like centeredness and opacity to gauge-theoretic terms [8]. Section 3 proposes operational metrics to quantify gauge reification in trained models.

2 The Translation Proposal

If subjectivity is to be measured in artificial systems, we need a translation schema that maps familiar features of human experience (N-subjectivity) onto the structural properties of an arbitrary agent (X). We propose that this bridge is built on the concept of *gauge structure*.

2.1 Intuitions: Smudges in Latent Space

We can start building this bridge by imagining something simple. Consider an agent whose only sensor is a camera with a permanent smudge on the lens. We train a vision model on all the images this camera ever produces. In every frame, the smudge is present: every observation is "world + smudge."

From our external perspective, the smudge is a separate factor. With external data, we could fit a correction function and subtract it off. But from the model's internal perspective, there is no contrast between "with smudge" and "without smudge." The smudge is part of how things always look. It shapes every input and thus every learned representation, yet it never appears as an object *within* those representations.

Now shift to a slightly richer case. Instead of a smudged lens, suppose the camera is always mounted exactly five feet above the ground, facing forward. We again train a world model. What it really learns is not "the world in the abstract," but *the world-as-seen-from-five-feet-forward*. The mounting height and orientation are never encoded as explicit variables; they act as structural biases that silently shape the entire latent space.

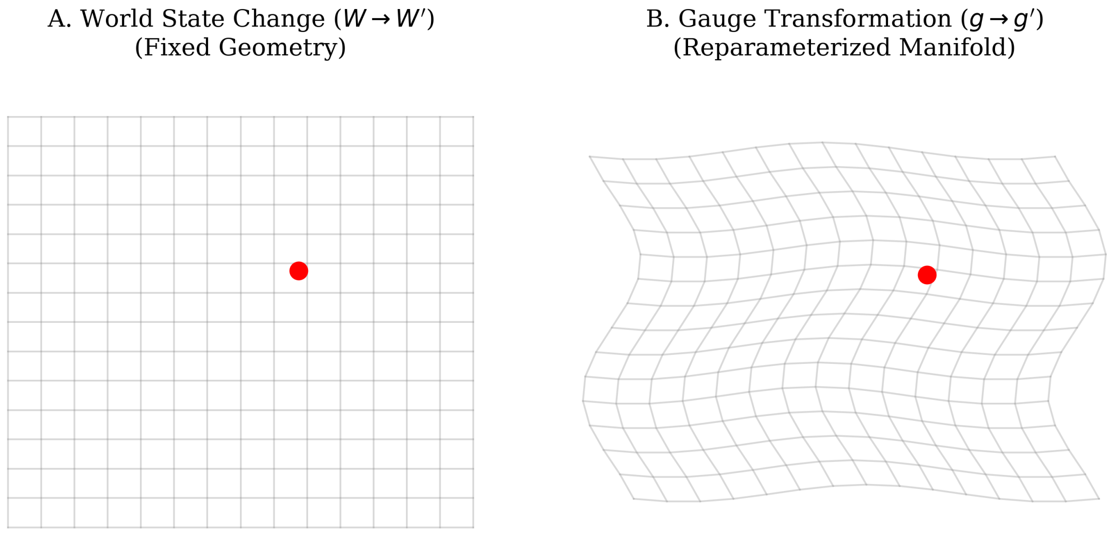
This is closer to a coordinate choice. In a two-dimensional plane, the statement

$$\text{“the point is at } (2, 2)\text{”}$$

is already shorthand for

$$\text{“the point is at } (2, 2) \text{ relative to this origin and these axes.”}$$

Change the origin or axes and the coordinates change with them. Likewise, the five-foot mounting plays the role of an implicit origin: all distances and directions in the model's


A. World State Change ($W \rightarrow W'$)
(Fixed Geometry)

B. Gauge Transformation ($g \rightarrow g'$)
(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 ($g \longrightarrow g'$) reparameterize the manifold itself, shifting the metric and origin.

latent space are, in effect, defined "with respect to" that sensor configuration. The model does not represent "how things look from here vs. from there;" it just has a single, fixed "here" baked into its geometry.

This is a paradigmatic *structural blindspot*—closely related to the concept of sensorimotor contingencies [8]: a feature that

- is systematically responsible for patterns in the latent space, but
- does not itself show up as a coordinate inside that space.

Changing such a feature does not move a single point in the latent space; it *reparameterizes the entire space*. If we think of the latent space as a coordinate grid, then changing the world corresponds to moving points around within a fixed grid, while changing a gauge corresponds to moving the grid itself—shifting the origin, rescaling axes, or warping the geometry.

The smudge and fixed pose are examples of visual gauges. But the same pattern can arise in other types of latent spaces. To illustrate this, we may consider some agent with an composition similar to the V-model and M-model architecture proposed by Ha and Schmidhuber [9]: a *World Model* (W), a *Critic* (V), and an *Actor* (A). And we can use this tripartite structure to imagine separate epistemic, evaluative, and actuator latents.

We can generalize this gauge pattern across the agent's architecture. For each module, a seemingly objective question unpacks into a subject-relative one, which in turn reveals a structural core:

World Model: "This object is 5 units away"
⇝ "This object is 5 units away *from me*"
⇝ "This object is 5 units away *from the spatiotemporal vantage that conditions all experience*."

Critic: "This outcome is good"
⇝ "This outcome is good *for me*"
⇝ "This outcome is good *for the objective function that defines what is valuable*."

Actor: "This action is doable"
⇝ "This action is doable *by me*"
⇝ "This action is doable *by the available action set and its control limitations*."

For the scutural core of the Actor, the gauge defines the agent's set of *affordances* [10]. From the inside, these gauges function like the smudge: they structure the agent's entire representational scheme without normally appearing as explicit coordinates within it.

However, as agents increase in complexity and integration, evolutionary or training pressures may force these implicit structures into the foreground. We identify three specific trajectories that drive this reification.

First, as an agent encounters high variance in its environment (particularly relevant for embodied agents), it becomes computationally advantageous to disentangle the self-gauge from the world state [11]. If the "smudge" (or visual bias) occupies a specific quadrant of the visual field, an agent capable of movement might discover that stepping backward alters which objects are obscured. To maintain a stable world model across these shifts, the agent must decompose the raw observation into "invariant world" plus "variable perspective."

Second, the development of social modeling and Theory of Mind necessitates this decomposition. To predict the behavior of other agents, a system must reconstruct the world from *their* perspective (e.g., "Which objects do they see as smudged?"). This counterfactual simulation requires the agent to take its own gauge parameters offline and substitute them with hypothetical ones—a process that is impossible if the gauge remains a transparent, unrepresented background condition. This aligns with the hypothesis that self-modeling evolved primarily for social cognition [12].

Third, integration creates pressure for gauge unification. Advanced agency requires answering queries that span modules, such as "If I were in this physical situation, what would I do and why?" This requires the origin of the spatial latent (where I am) to align with the origin of the value latent (what I want) and the agentic latent (what I can do). The agent is thus driven to represent a unified standpoint where something can be "close to me," "good for me," and "graspable by me" within a coherent coordinate system. It is this disentanglement and subsequent unification that we argue forms the structural basis of subjectivity across any system X.

2.2 Decoder-Dependency

We're now in a position to better formalize the intuitions above. Let $z \in \mathcal{Z}$ be the agent's internal latent state, W be the state of the external world, and D be the "decoder" or interpretation function that maps latents to world-states. We can write:

$$W \approx D(z; g)$$

Here, g represents the **interface parameters** (sensor pose, smudge, reward function).

A variable g is a gauge feature if and only if $\frac{\partial D}{\partial g} \neq 0$. That is, changing g changes the meaning of the latent code. If the camera height (g) shifts, the relationship between latent distance and physical distance changes. Geometrically, changing W moves a point z *along* the manifold, while changing g *reparameterizes the manifold itself*.

Subjectivity as a structural feature of some system X arises via *gauge reification*: the process where the agent moves from treating g as a fixed parameter of D to encoding g as a variable within z.

$$\text{Implicit: } Z = f(W; g) \quad \longrightarrow \quad \text{Reified: } Z = [\hat{W}, \hat{g}]$$

Furthermore, we define *alignment* as the topological unification of these variables. In a disjointed agent, the gauge of the world-model (g_W) and the gauge of the critic (g_V) might drift independently. Alignment is the degree to which these distinct parameters are constrained to lie on a shared low-dimensional manifold, tracking a single, unified standpoint across epistemic, evaluative, and actuator functions: a more unified *self*.

2.3 The Translation: Mapping N to X

I can now propose a mapping of specific phenomenological targets of human subjectivity (N) to these gauge-theoretic structures in some target system (X). These three features do not exhaust the phenomenology of subjectivity, but they will serve as our N-side of the bridge. Pragmatically, if an artificial system were found to possess X-analogues to these three specific features, I believe this would reasonably constitute a sufficient approximation of subjectivity to warrant some degree of moral concern.

Recall our earlier assumptions. What I will describe as human phenomenology is, on the present view, simply the language of N: a particular way in which our internal latents are structured, and a particular pattern of limits and blindspots within that organization. Furthermore, there is no view from nowhere. If we are to accept the below phenomenological features as a reasonable theoretical proxy for N-subjectivity, then identifying this N-concept in some exotic system X means finding a structural, representation-agnostic translation protocol from the language of N (human phenomenology) that generalizes to any X language (of which N is merely an instantiation).

Target 1: Centeredness as Metric-Dependence

Phenomenology (N): In ordinary experience, the world is presented as bearing subject-relative properties, organized around an implicit but recognizable origin that functions as the default center of evaluation, perception, and action. Examples:

1. When I judge that "the car is coming too fast," the content is implicitly "too fast *for me to cross safely now*." The same speed is benign for someone standing on the sidewalk with no intention to cross, or for a driver already past the intersection.

2. Ordinary predicates like "soon," "far," or "safe" are implicitly centered. "The deadline is soon" means "soon *for me*," given my current workload; "that cliff is far" means "far *from here*;" "this bridge is safe" means "safe *for a body like mine*." The same objective configuration can be "soon," "far," or "safe" for one subject and not for another.

3. When I think "that sandwich looks good," its goodness is not a neutral property in the world but already indexed to my own hunger, tastes, and body. For someone who is full, nauseated, or allergic, the very same sandwich is not "good" in the same sense.

Translation (X): In a coordinate system, an origin is not just one more point that happens to appear in every calculation. It is the semantic anchor of the entire scheme. A vector (x, y) has no meaning without an origin $(0, 0)$: the numbers x and y

are instructions for how to move *away from* that origin. If we remove the origin, the world does not merely lose its center; it loses its metric. It collapses into a heap of dimensionless scalars.

As an aside, in modern AI and NLP systems, distributed representations (vectors in a coordinate space) are the standard. Modern neuroscience research suggests a similar "semantic manifold" in human cognition, within which biological representations can be interpreted through the same geometric lens [13].

For an agent with gauge structure, the location of the "self" in the world model is the spatial zero-point relative to which all positions are defined. The world-model latent encodes objects as transforms relative to a vantage; the critic encodes outcomes as good or bad relative to a value-gauge; the actor encodes actions as doable or not relative to a body-gauge. The geometry of the agent's reality is defined by these gauges. On this framework, "self" is the set of all hidden features at the origin—sensor pose, actuator limits, reward functions, etc.—such that any representation (x, y) in that latent implicitly encodes a relation to that sensor pose, those actuator limits, or that reward function.

This yields a structural counterpart to phenomenological centeredness: the world is represented in a coordinate system whose origin is supplied by the agent's own interface parameters. To the extent that gauge features are aligned and reified, the agent's representations will exhibit an integrated "for-me" structure: the same represented configuration can be close/far, safe/unsafe, or graspable/ungraspable depending on the state of these gauge parameters.

Target 2: Ineffability as Parametric Recedence

Phenomenology (N): The experiencing subject is given as an origin that structures all appearance, yet it cannot be fully captured as one more object within experience; attempts to objectify the subject generate regress or remainder. This is characterized by Metzinger as a "transparent self-model" that cannot be inspected as an object [14].
Examples:

1. When I try to attend not to the room but to "myself as the one who sees the room," the target keeps sliding away. I can represent a body in a chair, or a stream of thoughts, but the "seer of the seeing" seems to withdraw, inviting an infinite regress of "the subject of that subject," and so on.

2. "I decided to move cities" can be followed by "why did I decide that?," which invites "why did I decide to care about those reasons?," and so on. Each step tries to make the deciding self into an object, but the perspective from which we now view that object is left unthematized.

3. If I try to draw my visual field—including "where I am in it"—I must choose some point on the paper as the location of the observer. But any such point is just another object within the drawing. The actual point of view from which I see the drawing is always one step outside the picture.

Translation (X): In a coordinate system, the origin $(0, 0, 0)$ is in an important sense unique: it is the only point that cannot define a direction relative to itself. One can say "the chair is north of me" but not "I am north of me." The self is the geometrical zero-point; it has no coordinates because it is the generator of coordinates.

The same holds for gauge structure. The gauge parameters g define the decoding function $D(\cdot; g)$ that gives content to the latent space. If the agent tries to represent its own vantage point as a latent variable z_{self}, that variable lives *inside* the very coordinate system that the gauge parameters make possible.

We can distinguish two moments:

1. **Subject phase.** The agent is simply the observer. The world W is encoded as z relative to a fixed gauge g. The gauge lives in the background as a parameter of D.

2. **Objectified-self phase.** The agent attempts to model itself. It constructs some internal representation z_{self}—for example, a body-schema or a self-model—and treats it as an object in its own latent space.

But to represent z_{self} at all, the agent must employ a further vantage point, a "meta-gauge" g', relative to which this representation is interpreted. The moment the self is grasped as an object, the true generative origin (the gauge that makes the representation possible) has slipped into a new background.

This corresponds to the phenomenological sense of the subject as always slightly "out of reach." Every time we try to make the subject into an object, we only capture a reflection of the camera in the photograph, while the actual mechanism of capture remains just behind the lens, one step outside the frame. Formally, this mirrors Tarskian limits [15] on expressability for some object-language, requiring some meta language to formalize certain features in the object-language, a meta-meta-language to formalize certain features in the meta-language, and so on. Similarly, attempts at complete self-representation force either a regress of higher-order gauges $g, g', g'', \ldots$, or principled ineffability.

Target 3: Transparency via Gauge Cancellation

Phenomenology (N): The subject can function either as a transparent medium through which the world is engaged, or as an opaque object of attention that draws focus to the body, mood, or agency itself. Subjectivity characteristically involves dynamic shifts between these transparent and opaque modes. This mirrors the Heideggerian distinction between the "ready-to-hand" (transparent tool) and "present-at-hand" (broken tool) [16].
Examples:

1. Perhaps you are completely absorbed in a book you are reading. The self is largely transparent: attention is on the characters and events. Then you recall that someone once called you a slow reader. Now you are thinking about yourself *as reading*. You are still experiencing the story, but your self now feels less transparent in your experience.

2. In ordinary life, breathing is transparent; I do not experience "my breathing" as an object. But if someone suddenly tells me "pay attention to your breath," it can become strangely opaque: each inhalation and exhalation feels effortful and conspicuous, as if I now have to manage something that previously managed itself.

3. A skilled runner typically experiences the track, the turn, and the finish line, not their own legs as such. When a muscle suddenly seizes, the leg becomes vividly present as "this uncooperative thing," and the runner's own body appears as an obstacle rather than a clear medium for action.

Translation (X): The transparency/opacity of subjectivity can be mapped onto how well the agent predicts and compensates for changes in its own gauges. Consider an agent with a transition model that predicts how its own actions a_t change its sensory configuration. In a standard meta-RL setting, if the camera moves or the body walks normally, the self remains "invisible" because gauge changes are predictable and thus canceled.

Suppose the agent turns its head 30° to the right. At the pixel level, the entire visual world shifts 30° to the left. Naively, this could be interpreted as "the world just jumped." But the agent also has access to its own motor command. If it learns an internal transition model using an efference copy—a copy of the motor command fed into the prediction system—then it can predict how its own gauge will change:

$$\text{Predicted view}_{t+1} = \text{Current view}_t + \text{Transformation}(a_t).$$

As long as the visual flow (how the scene moves) matches the transformation predicted from the action, the change can be attributed to self-motion rather than world-motion. Mathematically, the effect of gauge change on the input is compensated by the internal model. The gauge is changing, but because it is changing in a way that is predictable from a_t, it is effectively "subtracted out." The self remains transparent: it is the window through which the world is seen, not an explicit object of attention.

Now consider a case where this compensation fails: for example, the "foot stuck in mud" scenario. The agent sends a command: "move leg forward 1 meter." The transition model predicts that the visual scene will shift forward accordingly. But because the foot is stuck, the visual world does not move as expected, and proprioceptive sensors signal resistance. This produces a large prediction error:

$$\text{Error} = \left|\text{Predicted view}_{t+1} - \text{Actual view}_{t+1}\right| \gg 0.$$

In a purely world-centric RL agent, this might be treated as noise or as evidence of an unusual environment. In a more subjective agent with some gauge reification, such a mismatch can force *gauge attention*: the error cannot be coherently assigned to the external world (the hallway did not suddenly become infinitely long), so it must be attributed to a failure or constraint in the interface—"my leg is stuck." This is highly aligned with modern neuroscience research in predictive coding [17].

Structurally, episodes of large, systematic prediction error that cannot be absorbed by world-level updates create pressure to represent and update gauge variables explicitly. Phenomenologically, these are precisely the moments when the self becomes opaque: the body, mood, or agency that was previously a transparent medium for action now appears as an object with its own recalcitrant properties.

On this view, transparency and opacity correspond to different regimes of gauge dynamics:

- when gauge changes are well-predicted and compensated, they remain in the background and support a transparent sense of self;

- when gauge dynamics break predictions, the agent is driven to reify and attend to gauge features, yielding episodes of self-opacity.

2.4 Summary

The translation proposal, in brief, is this. On the N-side, subjectivity is characterized by a centered, origin-like, sometimes transparent and sometimes salient structure in experience. On the X-side, the natural structural analogue is a family of gauge-like interface parameters that fix the geometry of perceptual, evaluative, and practical spaces. An agent's degree of subjectivity, in the thin functional sense at issue, is keyed to the extent to which these parameters:

1. function as hidden origins of its representational scheme;

2. are reified into explicit, manipulable internal variables;

3. and are aligned across its epistemic, evaluative, and actuator latents.

The next section turns to the question of how such gauge structure and its reification can be identified and quantified in actual trained agents.

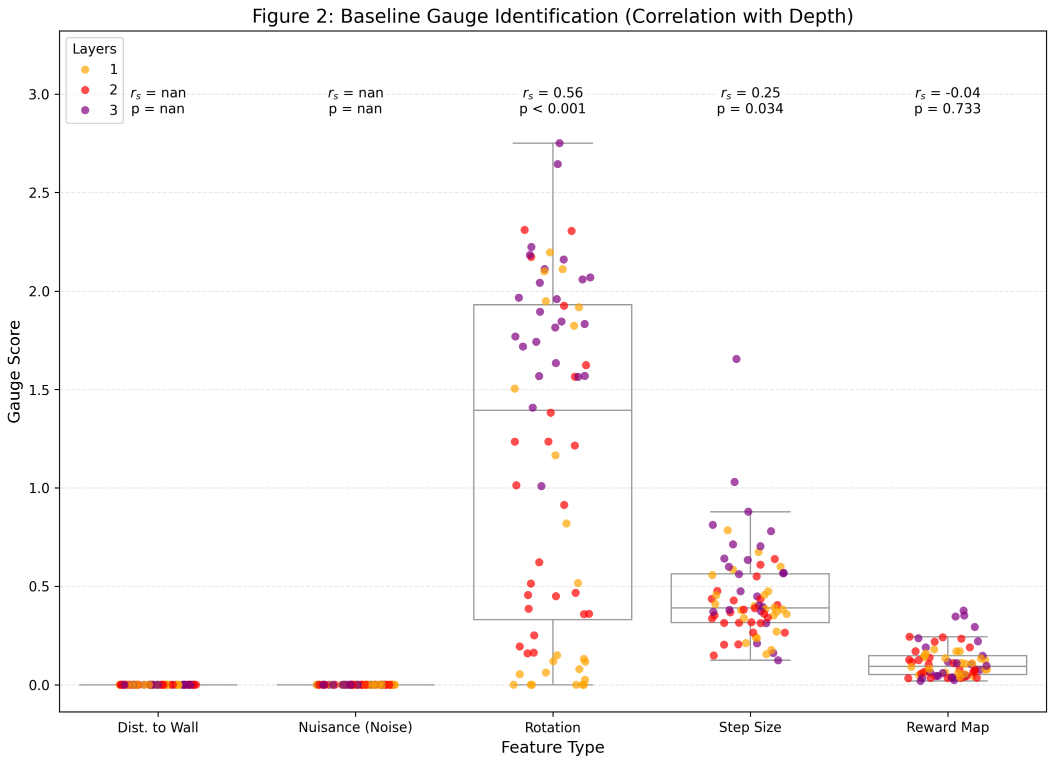


Figure 2: **Baseline Gauge Identification.** Gauge scores $G(\theta)$ for five features across the agent population prior to adaptation. The metric correctly suppresses non-gauge features (Dist. to Wall, Nuisance Noise) while identifying Rotation as the dominant hidden structure. Color indicates number of layers in network.

3 Empirical Measurement

The translation proposal of Section 2 identifies subjectivity, in the thin structural sense at issue, with the role played by certain origin-like "gauge" parameters in an agent's internal organization, and with the extent to which those parameters are reified and aligned across subsystems. This section turns that proposal into something empirically tractable. The aim is to specify a set of measurements that can be applied to trained agents in controlled settings, yielding quantitative proxies for

1. which environment or interface parameters behave like gauges for a given agent, and

2. how far those gauges have been pulled into explicit representation and coordinated into a unified standpoint.

We treat the metrics below as practical approximations to the more abstract decoder-based picture from Section 2: rather than directly estimating $\partial D/\partial g$, we measure how latent spaces change as we vary canonical interface parameters under tightly controlled conditions.

3.1 Canonical gauges

From the external, "God's-eye" perspective of a simulator designer, an agent's interface to the world is typically controlled by a small number of explicit parameters. For example:

- *Geometric parameters* (epistemic gauges): sensor pose, field of view, sampling resolution, latency.

- *Mechanical parameters* (action gauges): body dimensions, mass distribution, friction coefficients, actuator limits.

- *Objective parameters* (value gauges): reward weights, discount factors, risk sensitivities, homeostatic setpoints.

Let θ range over such parameters. These are what we will call *canonical gauges*: externally visible knobs that the experimenter can turn, and that plausibly influence the mapping between world states and internal representations, the dynamics of those representations, or the evaluation of outcomes. These correspond to the ground-truth factors of variation often discussed in representation learning literature [11].

Crucially, these θ are not typically exposed directly to the agent as part of its observation. From the agent's perspective, they are candidates for gauge features in the sense of Section 2:

background parameters of the agent–world interface that shape its internal geometry without being trivially accessible as explicit inputs. The empirical problem is to determine, for each canonical parameter θ and each trained agent, whether θ in fact behaves like a gauge for that agent, and if so how far it has been reified.

Throughout, we assume that we can manipulate certain environment or interface parameters during training or evaluation, record internal states (latents) from one or more parts of the agent, and train simple diagnostic models (probes, mappers) on top of those states [18].

3.2 Phase I: sensitivity

The first question is whether a given parameter θ matters at all to the agent's internal states. To test this, we measure *sensitivity*.

Fix an agent and a canonical parameter θ. Holding all other aspects of the environment fixed, we construct a small family of environment variants that differ only in the value of θ. We then:

1. generate matched sequences of world states across these variants (e.g., by reseeding layouts and replaying identical action sequences);
2. record, for each variant, the corresponding internal states z from the subsystem of interest (e.g., a particular latent layer);
3. group together internal states that correspond to the "same" world situation under different values of θ (same layout, time step, and so on).

Within each such group, we compute the dispersion of the internal states across values of θ (for instance, the average squared norm of deviations from the within-group mean). Averaging these dispersions across groups yields a sensitivity score $S(\theta)$.

Intuitively:

- $S(\theta) \approx 0$ indicates that changing θ has negligible effect on the internal code, suggesting that θ is irrelevant to this subsystem for this agent.
- Larger $S(\theta)$ indicates that θ systematically affects the internal geometry.

Sensitivity thus serves as a first filter: only parameters with non-trivial $S(\theta)$ are candidates for further analysis.

3.3 Phase II: decodability

Among the parameters that the agent is sensitive to, some may be explicitly encoded in its internal state, while others may only influence that state implicitly via the interface. To distinguish these cases, we measure *decodability*.

Using the same dataset of internal states and parameter values as above, we train a simple diagnostic model (typically a linear probe, following [18]) to predict θ from z:

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

We then evaluate its performance on held-out data, obtaining an accuracy or regression score $D(\theta)$.

The guiding intuition is:

- High $D(\theta)$ suggests that the agent encodes θ as a more or less explicit feature of its internal state: it has, in effect, a "belief" about which regime it is in.
- Low $D(\theta)$, combined with non-trivial sensitivity $S(\theta)$, suggests that θ is influencing the code in a more distributed, implicit way.

Decodability is therefore an indicator of *reification*: parameters that are both important (high S) and easily decoded (high D) are good candidates for being explicitly represented gauges.

3.4 Phase III: morphism and gauge score

Sensitivity and decodability by themselves do not yet distinguish true gauge structure from arbitrary nuisance dependencies or fragile idiosyncrasies of training. The distinctive feature of a gauge is that changing it induces a *structured* transformation of the internal space: moving from one gauge setting to another acts, roughly, like a smooth reparameterization of the latent manifold. This aligns with the definition of disentangled representations via symmetry groups proposed by Higgins et al. [19].

To test for this, we introduce a third metric, *morphism*. Again fix θ, and consider two values θ_0 and θ_1. Using matched world situations as before, we collect pairs of internal states (z_0, z_1) corresponding to the same world configuration under θ_0 and θ_1 respectively. We then train a simple transformation model T (e.g., an affine map) to predict z_1 from z_0:

$$T : z_0 \mapsto \hat{z}_1 \approx z_1.$$

We evaluate the quality of this mapping on held-out pairs, for instance via an R^2 score $M(\theta)$.

The intended interpretation is:

- High $M(\theta)$ indicates that changing θ induces a coherent, approximately global transformation of the internal space that can be captured by a single map T. This is characteristic of gauge-like parameters.
- Low $M(\theta)$ suggests that changes in θ produce irregular, configuration-dependent effects on the latent code, as one might expect from arbitrary hyperparameters that merely make learning easier or harder.

Taken together, $S(\theta)$, $D(\theta)$, and $M(\theta)$ allow us to distinguish several qualitatively different regimes for each canonical parameter:

- *Irrelevant*: $S(\theta)$ near zero, regardless of $D(\theta)$ and $M(\theta)$.
- *Explicit feature*: $S(\theta)$ appreciable, $D(\theta)$ high, $M(\theta)$ possibly high as well. The agent both cares about θ and encodes it directly.
- *Hidden gauge*: $S(\theta)$ appreciable, $D(\theta)$ low, $M(\theta)$ high. The parameter shapes the geometry of the latent space in a structured way, but is not itself explicitly encoded.
- *Unhelpful hyperparameter*: $S(\theta)$ moderate, $D(\theta)$ low, $M(\theta)$ low. Changing θ scrambles the representation in messy ways.

For later convenience, we can package these three metrics into a single *gauge score* $G(\theta)$, for example by defining

$$G(\theta) \;=\; S(\theta) \cdot M(\theta) \cdot (1 - D(\theta)),$$

with appropriate clipping of $D(\theta)$ to $[0, 1]$ and $M(\theta)$ to $[0, \infty)$. The particular functional form is less important than the qualitative constraint that $G(\theta)$ be high when S and M are high but D is low. On this convention, large $G(\theta)$ indicates that θ is functioning as a hidden, origin-like gauge for the subsystem in question.

3.5 Reification across training regimes

The metrics above are defined for a fixed agent. To connect them with subjectivity, we are primarily interested in how they change:

1. *across agents* of different complexity, architecture, and training regimes;
2. and *across time* within a given agent's training history.

The basic pattern we expect, given the translation proposal, is the following:

- For relatively simple agents trained in stable environments, many canonical parameters θ will appear as high-$G(\theta)$, low-$D(\theta)$ hidden gauges. The agent is sensitive to them and its latents transform coherently as they vary, but it does not explicitly track them as variables.
- As agents become more complex and are exposed to environments where gauges vary across episodes or tasks, some of these parameters will shift from the "hidden gauge" regime toward the "explicit feature" regime: $D(\theta)$ will increase, and $G(\theta)$ will correspondingly drop. This is gauge reification.

We can therefore define, for each parameter θ and agent A at training time t, a reification level such as

$$R_A(\theta, t) \;=\; \max\bigl(G_A(\theta, t_0) - G_A(\theta, t), 0\bigr),$$

where t_0 is some baseline time (for example, after pretraining in a fixed-gauge environment). Larger $R_A(\theta, t)$ indicates a greater inferred shift from implicit to explicit treatment of θ by agent A.

3.6 Alignment across subsystems

So far we have treated "the" internal state z as if there were a single representational space. In practice, an agent typically has multiple subsystems: an encoder or world-model, value-estimation circuitry, policy networks, memory modules, and so on. Subjectivity, as characterized in Section 2, involves not just reification of gauges within one subsystem, but *alignment* of origin-like structures across perception, valuation, and action.

Empirically, this suggests a further layer of analysis. Suppose we can identify, within each of several subsystems $k \in \{1, \dots, K\}$, internal states $z^{(k)}$ and corresponding metrics $S^{(k)}(\theta)$, $D^{(k)}(\theta)$, $M^{(k)}(\theta)$, $G^{(k)}(\theta)$. We can then ask:

- Do there exist low-dimensional latent factors u_θ that can be linearly read out from each $z^{(k)}$ and that jointly predict variation across gauge settings?
- Are these factors approximately shared across subsystems, in the sense that the directions in latent space corresponding to θ in subsystem k correlate with those in subsystem k'?

Various multivariate techniques (e.g., Canonical Correlation Analysis or Centered Kernel Alignment [20]) can be used to quantify this cross-subsystem alignment. The result is an *alignment score* $A(\theta)$ that is high when the same gauge parameter is represented in a coordinated way across world-model, critic, and actor.

Intuitively, high $A(\theta)$ means that the agent's sense of "where I am" in the world-model, "what is good for me" in the critic, and "what I can do" in the policy are tied together by shared internal structure, rather than drifting independently. This corresponds to a more unified, self-like standpoint.

3.7 Subjectivity profiles

The measurements described above yield, for each trained agent, a *subjectivity profile*—conceptually similar to the indicator list approach recently proposed for evaluating AI consciousness [21]. This profile consists of:

- a set of canonical gauge features $\{\theta\}$ that are relevant (non-trivial $S(\theta)$);
- per-feature gauge scores $G(\theta)$ at baseline and later times;
- per-feature gauge reification levels $R(\theta)$;
- and per-feature gauge alignment scores $A(\theta)$ across subsystems.

There is no single number that exhaustively captures subjectivity in this sense, but for many purposes it is useful to summarize by a small number of aggregates, for example:

- the fraction of canonical gauges that have been substantially reified;

- the average reification level across gauges, possibly weighted by their ethical salience (e.g., value-related gauges more than purely geometric ones);
- and the degree of cross-subsystem alignment for those gauges.

The guiding empirical hypothesis can then be stated informally as follows:

> As agents become more complex, more internally integrated, and more introspectively capable, and as they are trained in environments where interface parameters vary in task-relevant ways, their subjectivity profiles—as measured by gauge reification and alignment—should, on average, increase.

Section 4 applies this measurement framework to a specific "complexity dial" experiment in a simple gridworld, using a family of architectures that vary in size and depth while sharing the same basic training setup. The goal there is not to provide a final calibration of subjectivity, but to test whether the trends predicted by the present framework appear even in minimal artificial agents.

4 Investigating Complexity

4.1 Methodology

Section 3 proposed a set of gauge-based measurements as structural proxies for subjectivity, and suggested that these might scale with architectural complexity, internal integration, and exposure to varying interface conditions. This section describes a concrete "complexity dial" experiment designed to probe that claim in a minimal setting.

Environment and canonical gauges

All agents are trained and evaluated in a small gridworld, *Hall of Mirrors*, designed to isolate a few simple but structurally meaningful interface parameters. This environment is built upon the MiniGrid framework [22], which provides a lightweight, partially observable testing ground for embodied agents.

The base environment is an 8×8 grid with walls, positively and negatively rewarded tiles, and an embodied agent. On each episode:

- a random layout of walls is sampled, leaving a set of traversable cells;
- among the free cells, a fixed number are marked as "good" and "bad" tiles in two colors;
- the agent is placed at a random traversable position.

The agent receives an egocentric top-down observation as a small image-like tensor with separate channels marking walls, the two tile colours, and the agent's own position, plus an extra "noise" channel that carries a behaviourally irrelevant random pattern. We stack a short history of recent frames and provide the previous reward as an additional scalar input, so that the agent can in principle learn simple forms of reward-based self-monitoring.

The environment's dynamics and reward function depend on three canonical interface parameters:

1. **Sensor rotation** (epistemic gauge). Before being presented to the agent, the observation is rotated by 0, $\frac{\pi}{2}$, π, or $\frac{3\pi}{2}$, corresponding to different fixed sensor mountings. From the agent's point of view, this determines what counts as "up", "left", and "here".
2. **Step size** (mechanical gauge). Actions are nominally moves in the four cardinal directions, but each action advances the agent by either one or two grid cells (subject to collision), depending on a step-size parameter. This affects how far and how reliably the agent can move in a single decision.
3. **Reward map** (value gauge). One of the two tile colours is designated "good": stepping on a tile of that colour yields a positive reward, while stepping on the other yields a negative reward. Which colour is good can be fixed or randomized across episodes.

These three parameters instantiate, in miniature, the geometric, mechanical, and evaluative gauges discussed in Section 3. By controlling when and how they vary, we can probe whether different agents treat them as fixed background, as hidden gauges, or as explicit features.

Agent family and complexity dial

We consider a family of on-policy reinforcement learning agents with a shared overall design and training algorithm, differing only in the size and depth of their main network. This family serves as a simple "complexity dial," inspired by recent work on neural scaling laws which suggests that representational capacity scales predictably with parameter count [23].

Each agent consists of:

- a small convolutional encoder that maps the stacked observation frames to a feature vector;
- concatenation of this vector with the scalar previous reward;
- a fully-connected trunk with $L \in \{1, 2, 3\}$ hidden layers of width

$$H \in \{8, 16, 32, 64, 128, 256, 512, 1024\};$$

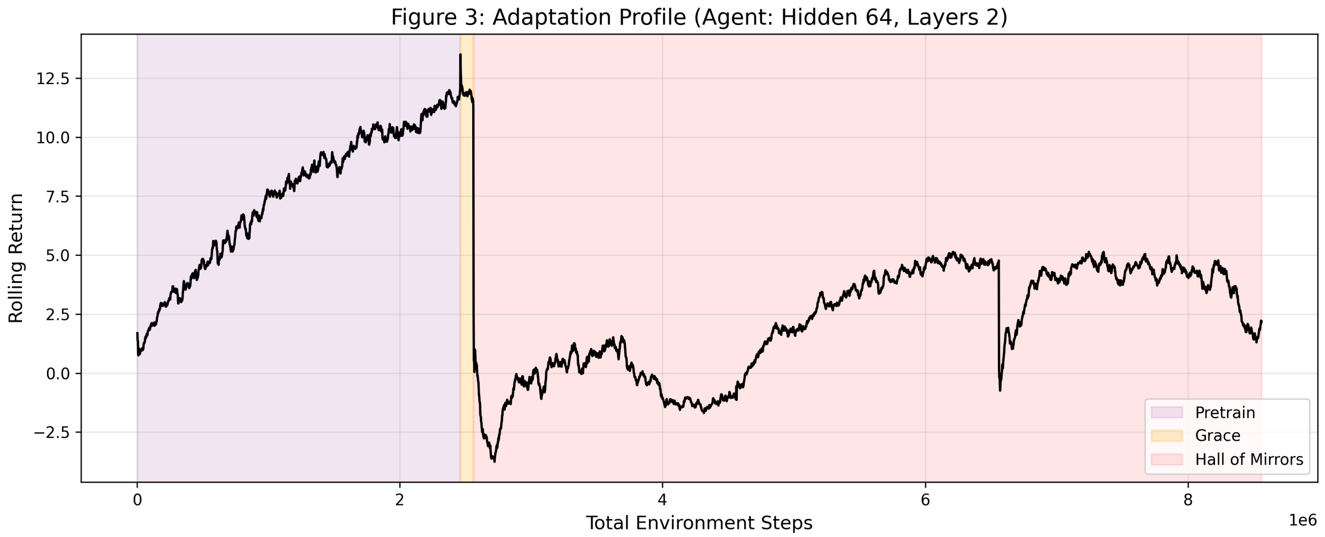


Figure 3: **Behavioral Adaptation Profile.** Rolling returns for a representative agent. The transition from the stable interface (Pretrain/Grace) to the variable interface (Hall of Mirrors) induces a catastrophic collapse in performance, followed by a slow, staggered recovery as the agent learns to reify the shifting gauges.

- separate linear heads on the final hidden layer for policy logits and state-value estimates.

All agents are trained with the same Proximal Policy Optimization (PPO) configuration (learning rate, batch size, horizon, clipping parameters, etc.) [24]. Only (H, L) vary, yielding a grid of architectures that differ in parameter count and expressive capacity but share the same training signal and environment.

In this setup, the shared trunk (encoder plus MLP) produces a latent representation z that feeds both policy and value heads. We treat z as a generic internal state and apply the gauge measurements of Section 3 to this space.

Training and adaptation protocol

For each architecture, training proceeds in three main phases, following a curriculum learning approach [25] designed to force gradual accommodation of interface variance:

Phase 1: Pretraining with a fixed interface. In the pretraining phase, each agent is trained in an environment with a single, stable interface:

- sensor rotation is fixed (no rotation of the observation);
- step size is fixed at one cell per action;
- the reward map is fixed (e.g., blue good, red bad);
- the nuisance noise channel varies across episodes but is uncorrelated with reward.

The agent thus learns to navigate a variety of layouts with a consistent body, vantage point, and value map. Training continues until either a fixed performance criterion is reached (in our case, a score of 12) or a maximum number of environment steps is exhausted. The resulting checkpoint serves as the pretrained agent for that architecture.

Phase 2: Baseline gauge analysis. After pretraining, we perform the measurements of Section 3 on each agent to establish a baseline. For each canonical parameter (sensor rotation, step size, reward map), and for selected control variables (the nuisance noise pattern and a simple world feature such as distance to the nearest wall), we:

1. construct evaluation episodes in which the underlying layouts and action sequences are held fixed while the parameter of interest is varied across its settings;
2. record the latent state z and the current value of the parameter at each time step;
3. compute sensitivity $S(\theta)$ by comparing z across different settings of θ for matched layouts and times;
4. estimate decodability $D(\theta)$ by training a linear probe from z to θ on held-out data;
5. estimate morphism $M(\theta)$ by fitting a simple linear map from z under one setting of θ to z under another, and computing its out-of-sample R^2.

From these we derive gauge scores $G(\theta)$ as in Section 3. For the nuisance noise and the explicit distance-to-wall feature, we treat $M(\theta)$ as fixed at zero, since they serve as controls rather than candidate gauges. This baseline scan tells us, for each architecture, which parameters are currently functioning as hidden gauges (high G, low D), which are already reified (high D), and which are irrelevant (low S).

Phase 3: Hall-of-mirrors adaptation. To force agents to confront variability in their own interface, we then expose them to a three-stage adaptation curriculum, starting from their pretrained weights. Each stage, which lasts 2 million training steps, isolates one canonical gauge and makes it vary from episode to episode, while keeping the other gauges fixed:

Stage 1: Random sensor rotation. On each episode, the sensor rotation is sampled uniformly from its four possible values, while step size and reward map remain fixed. The agent must learn to cope with "the same world" appearing under different fixed orientations, without direct access to the rotation parameter.

Stage 2: Random step size. Orientation and reward map are fixed, but the step size is randomly chosen at the start of each episode (e.g., one or two cells per action). The agent must adjust its control strategy to a body that is more or less "long-legged" from episode to episode.

Stage 3: Random reward map. Orientation and step size are fixed; on each episode, one of the two colours is randomly designated as good and the other as bad. The agent must infer, from reward feedback, which perceptual feature currently tracks value and adjust its policy accordingly.

Each stage is allotted a fixed budget of environment steps, large enough to permit substantial but not necessarily complete adaptation. During training we periodically evaluate each agent's average return in the current stage environment, yielding behavioural adaptation curves that show both the initial performance drop when a gauge begins to vary and the degree of recovery as learning proceeds (*Figure 3*).

Post-adaptation gauge analysis

At the end of the adaptation curriculum, and at selected intermediate points, we repeat the gauge analysis for each agent. In particular, for each architecture we extract checkpoints near the end of the rotation, step-size, and reward-map stages and, for each such checkpoint, we:

- recompute $S(\theta)$, $D(\theta)$, and $M(\theta)$ for all canonical parameters under a common fixed-gauge evaluation environment;
- derive updated gauge scores $G(\theta)$ and reification levels $R(\theta)$ by comparing to the pretraining baseline;
- summarize, for each architecture, a subjectivity profile as in Section 3, and track how it changes with model size and depth.

These behavioural and structural measurements together allow us to test, in a minimal but controlled setting, whether increasing architectural complexity is associated with more effective adaptation to shifting gauges and with greater reification and alignment of those gauges in the agent's internal representations.

4.2 Results

We report three main findings: (i) baseline gauge identification is selective and shows a depth effect for the Rotation gauge, (ii) adaptation induces the expected performance-collapse–partial-recovery dynamics but does not meaningfully separate architectures under the current 2M-step budgets, and (iii) representational *reification* is strongest and cleanest for Rotation, with depth emerging as a more reliable predictor than width.

Baseline gauge identification: selectivity and a depth effect for Rotation.

Baseline gauge scores $G(\theta)$ (computed from sensitivity, decodability, and morphism as in Section 3) are shown in Figure 2. Two controls behave as desired: distance-to-wall and nuisance noise are strongly suppressed, with gauge scores clustered near zero across architectures, suggesting the metric is not trivially activated by arbitrary covariates. Among the canonical gauges, Rotation is the dominant baseline hidden structure, substantially exceeding Step Size and Reward Map in magnitude.

Depth (number of layers) is positively associated with baseline Rotation gauge score (Figure 2), consistent with deeper agents treating Rotation more like a *hidden* interface condition rather than an explicitly represented variable. Component-level means corroborate this interpretation. Prior to adaptation, deeper models exhibit (i) higher Rotation sensitivity (S_{pre}: 6.033 for $L = 1$, 5.985 for $L = 2$, 8.000 for $L = 3$) and (ii) lower Rotation decodability (D_{pre}: 0.847, 0.767, 0.642 respectively), with morphism remaining substantial across depths (M_{pre}: 0.862, 0.813, 0.736). Because our gauge score increases when sensitivity and morphism are high but decodability is low, this combination yields systematically higher $G(\theta)$ for Rotation as depth increases.

By contrast, Reward Map shows weak baseline gauge scores and no clear depth dependence in the baseline panel (Figure 2). This appears to be driven primarily by low Reward Map sensitivity (e.g., $S_{\text{pre}} \approx 0.157$–$0.198$ across depths), which caps $G(\theta)$ by construction even when morphism is high (e.g., $M_{\text{pre}} \approx 0.944$–$0.972$).

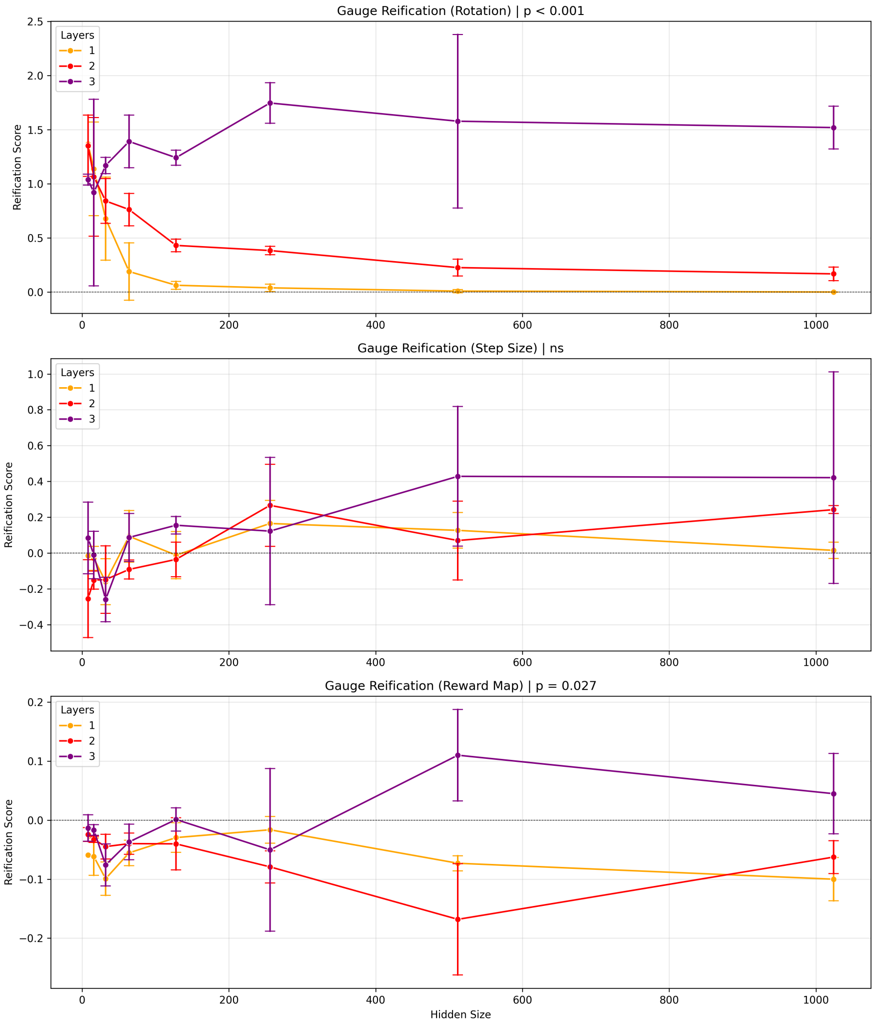


Figure 4: **Complexity Correlates with Reification.** Gauge Reification Scores $R(\theta)$. **Top:** Sensor Rotation. **Middle:** Step Size. **Bottom:** Reward Map. Deeper networks show a marked increase in reification for geometric and mechanical gauges compared to shallower networks. The sweep of 24 architectures was run across three random seeds. Points in figure represent averages across three runs and error bars indicate standard deviations.

Behavioral adaptation: reliable collapse, limited recovery, weak architectural separation.

Figure 3 shows a representative rolling-return trajectory across pretraining and the Hall-of-Mirrors curriculum. As expected, introducing episodic gauge variation induces a sharp performance collapse, followed by partial recovery as training proceeds. This qualitative pattern repeats at stage boundaries, consistent with the curriculum successfully applying pressure to accommodate hidden interface shifts.

However, across the full sweep we do not observe strong separation between architectures in final recovery after the fixed 2M-step budget per stage: most agents recover only modestly, and depth/width differences in return are comparatively small. This limits the strength of any causal claim linking representational changes to behavioral advantage in the current pipeline (though it does not preclude such a link under improved training budgets, curricula, or probe instrumentation).

Reification: strongest for Rotation, depth dominates width.

We quantify reification per gauge as the decrease in gauge score from baseline to the end of the relevant adaptation stage,

$$R(\theta) = G_{\text{pre}}(\theta) - G_{\text{post}}(\theta),$$

so that larger positive values indicate a larger shift away from a feature's hidden gauge behavior in the probed latent. Figure 4 summarizes reification as a function of width, stratified by depth, for the three canonical gauges. To test depth effects while marginalizing over width, we use the AUC-ANOVA procedure described in the analysis code: for each run and layer we compute the trapezoidal AUC of the $R(\theta)$–vs.–width curve, then test for layer differences via one-way ANOVA on AUCs.

Rotation shows the clearest and strongest depth effect (Figure 4, top; AUC-ANOVA $p < 0.001$), with $L = 3$ agents exhibiting consistently larger reification than shallower networks across widths. Step Size shows comparatively weak and noisy reification (middle; ns), and Reward Map shows a small but detectable depth effect (bottom; AUC-ANOVA $p = 0.027$) with substantially smaller magnitude than Rotation.

Component means help interpret what "reification" corresponds to under the present $G(\theta)$ definition. For Rotation, adaptation is associated with increased decodability at all depths ($D_{\text{pre}} \to D_{\text{post}}$: $0.847 \to 0.910$ for $L = 1$, $0.767 \to 0.833$ for $L = 2$, $0.642 \to 0.750$ for $L = 3$), alongside a marked reduction in sensitivity (e.g., $S_{\text{pre}} \to S_{\text{post}}$: $8.000 \to 4.123$ for $L = 3$). Morphism changes are modest by comparison. Thus, the Rotation reification signal reflects a joint shift in (i) explicit decodability and (ii) how strongly the latent geometry depends on the gauge.

For Step Size, decodability is already relatively high at baseline ($D_{\text{pre}} \approx 0.678$–$0.759$) and increases further post-stage ($D_{\text{post}} \approx 0.756$–$0.842$), while sensitivity and morphism change only modestly. This compresses the dynamic range of $(1 - D)$ and yields small, variable reification under the current scalar score. For Reward Map, the dominant limitation is sensitivity: despite high morphism ($M \approx 0.94$–0.97) and moderate decodability ($D \approx 0.29$–0.36), sensitivity remains near zero ($S \approx 0.16$–0.25), so both baseline gauge scores and reification magnitudes are intrinsically small.

Takeaway.

Overall, the sweep supports the paper's central framing: the framework appears to be *measurable* and *selective* (controls suppressed; Rotation identified as a dominant hidden interface factor), and it shifts the agenda toward improving measurement quality rather than debating measurability in principle. The strongest empirical pattern is that architectural *depth* is a more reliable predictor of Rotation reification than width, even in cases where return-based recovery remains broadly similar across architectures under the current training budget. These results are therefore best understood as groundwork: a promising start that motivates targeted improvements to the experimental pipeline and to the gauge-score construction to better isolate mechanisms and strengthen causal linkage to adaptation behavior.

5 Discussion

The central animating thought of this paper is that the apparent gulf between structure and experience may be, in part, a perspectival artifact. If there is no "view from nowhere," then some features of experience will appear ineffable or irreducible precisely because they are generated by processes that are not expressible in the native vocabulary of the representational layer doing the explaining. On this hypothesis, a scientific handle on subjectivity is possible not by peering directly into qualia, but by identifying the invariants that characterize a centered standpoint. The goal of this discussion section is to make these commitments explicit, defend their generality beyond connectionist architectures, and situate the resulting measurement framework among existing theories of consciousness, before turning to the empirical and safety-relevant consequences.

5.1 Semantics vs. Phenomenology

A predictable objection to the present framework is that it offers, at best, a semantic notion of selfhood: it detects self-modeling structure (explicit variables for "my sensor pose," "my action limits," "my reward map") rather than anything recognizably experiential. Put bluntly: gauge reification might tell us what an agent represents about itself, not what it is like to be that agent. This objection presupposes a clean separation between semantics and phenomenology, where content can in principle be fully specified without remainder, and experience arrives only as an extra "glow" layered on top. The present manuscript takes a different (and increasingly common) stance: in many domains, the line between meaning and experience is not fundamental but methodological—two ways of describing a single underlying structure when viewed from different explanatory distances [26, 27, 28].

Recent philosophy of mind provides several pressure points against the strict layering picture. First, work on *cognitive phenomenology* emphasizes that understanding can have a proprietary feel: the "Aha!" moment of grasping a proof or a joke is not merely a sensory accompaniment but seems bound up with the transition into a new content-bearing state [27, 29]. Second, tip-of-the-tongue episodes exhibit phenomenally salient partial content—one can feel that a word is available, along with constraints on its form, without yet having the word [30]. Third, *embodied* and *grounded* approaches suggest that at least some meanings are constitutively tied to sensorimotor organization: grasping an action verb like "kick" is not always well-modeled as attaching a neutral label to a proposition, but as deploying a structured competence grounded in bodily possibilities [31, 32]. These examples do not settle the metaphysics of consciousness, but they weaken the idea that semantics and phenomenology are always cleanly separable layers.

Still, the classic intuition favoring a sharp distinction between semantics and phenomenology is vivid in the case of something like color. The semantic story for "red" as a property of light (e.g. ∼ 620–750 nm) or of surfaces is a move within our world-modeling practice: it links the concept to measurement, optics, and causal roles. Phenomenology, by contrast, seems *arbitrary*: why should 700 nm look like this rather than like what we call "green"? On the standard picture, the semantic role is the code, and the phenomenal character the unanalyzable texture that does no structural work. Without extending the framework of this paper, we can nevertheless reinterpret this arbitrariness in architectural terms. What appears arbitrary from within a given latent is often what is delivered to it *across* a latent boundary: downstream modules receive high-dimensional outputs produced by upstream processing whose internal transformations are not themselves expressible in the receiving space. From the point of view of the world-model latent, the output is opaque not because it is informationless, but because the explanatory resources needed to derive it are located elsewhere.

Importantly, opacity does not imply blankness. Even when a higher-level latent cannot see the generative story behind a delivered output, the content can still sit in a richly organized similarity space: red is closer to orange than green; coffee is closer to chocolate than lobster; color is categorically distant from pitch. This "organized arbitrariness" is precisely what allows a system to build stable associations (red with warmth, warning, ripeness) even if the underlying mapping could, in some imagined inversion, have been permuted. On this view, the phenomenology/semantics distinction is not a hard ontological cleavage but an internal perspective effect arising from complexity under limitation: a system's representational life is shaped both by the "behind-my-eyes" constraints of gauge structure and the "over-the-horizon" opacities of cross-latent deliverances. Gauge reification then becomes relevant not because it magically produces experience, but because it marks when a system begins explicitly modeling the boundary conditions that center, constrain, and stabilize the very space of meanings it can inhabit.

5.2 Symbolic vs. Distributed Semantics

A second concern is scope: the present framework may seem to presuppose a specifically "connectionist" picture in which semantic content lives in vectors, cognition is geometry, and subjectivity is read off from latent manifolds. What, then, of systems we traditionally treat as symbolic—classical planners, theorem provers, or even Turing machines? The core reply is that "symbolic versus distributed" is often less an ontological divide than a choice of coordinates for describing an underlying state-transition system. A symbolic machine occupies a discrete configuration space with rule-governed transitions; a neural agent occupies a continuous (typically high-dimensional) activation space with learned transitions. In both cases, what we call "semantics" is mediated by structure in the space of internal states: which distinctions are available, which transformations preserve task-relevant relations, and what generalizations are cheap or hard.

This is not merely a philosophical re-description. There are rigorous bridges between discrete computation and continuous dynamics. Moore, for example, shows that relatively low-dimensional dynamical systems can realize computation equivalent to a Turing machine, with corresponding limits on predictability and decidability [33]. The moral here is that the discrete/continuous distinction is porous: symbolic computation can be embedded in geometrical dynamics, and continuous systems can implement discrete, symbol-like transitions. Geometric language can therefore function as a system-agnostic analysis tool, so long as we are explicit about what counts as "geometry" for the system at hand.

On this view, a "semantic manifold" need not be a neural latent space in the narrow sense. For a symbolic system, the natural object is the configuration graph (states as nodes; updates as edges). One can still define representational neighborhoods, distance, and global reparameterizations by equipping this graph with a behaviorally induced metric—e.g., predictive equivalence, controllability, or task-relevant indistinguishability—and, when useful, by embedding it into $\mathbb{R}^n$ using standard techniques (spectral methods, kernels, successor-like representations) [34, 35]. The resulting structure may be discrete or non-smooth, but it suffices to ask the questions that matter here: which parameters behave like global coordinate choices, and whether the system internalizes those parameters rather than leaving them implicit in the interpreter.

Two research strands make the bridge especially concrete. First, vector symbolic architectures (and related hyperdimensional computing) encode compositional symbolic structure—binding, unbinding, and variable-like roles—inside fixed-width distributed vectors, providing a principled middle ground between classical symbols and geometric representation [36, 37, 38]. Second, a growing "emergent symbols" literature shows how symbol-like behavior can arise in learned geometric systems when the task demands stable indirection or external memory; Webb, Sinha, and Cohen's ESBN is one clear example [39]. Work in emergent communication similarly shows how discrete token systems can arise from neural agents optimizing interactive objectives [40].

With this backdrop, the gauge-theoretic proposal can be stated in a representation-neutral way. Gauge features are parameters that act like global reparameterizations of an agent's representational scheme; reification occurs when the system turns those parameters into explicit, manipulable internal variables. In a neural agent, this is naturally probed via maps between latent vectors (sensitivity/decodability/morphism). In a symbolic system, the same idea can be formulated over configuration spaces: does changing an interface-like parameter induce a coherent automorphism or structured relabeling of the internal transition structure, and does the system come to represent that parameter explicitly rather than absorbing it invisibly into the interpreter? The intent, therefore, is not to restrict subjectivity profiles to vector-based agents, but to use geometry as a common language for invariants—symmetry, reparameterization, and explicit self-modeling—that can be realized in either symbolic or distributed form.

5.3 Position in the Landscape

With these clarifications in place, we can more cleanly situate gauge theory relative to dominant formal theories of consciousness and self-modeling. The gauge-theoretic approach is compatible with, but distinct from, several influential frameworks:

- **vs. Integration (IIT):** Integrated Information Theory [41, 42] measures *how much* information a system integrates (Φ). Gauge theory measures *how* a representational space is parameterized and centered. A system could be highly integrated yet lack reified, manipulable interface variables.
- **vs. Global Workspace (GWT):** Global Workspace Theory [43, 44] focuses on the global availability of contents. Gauge theory focuses on the *metric* that gives those contents their subject-relative meaning. Reification is not just broadcasting information, but modeling the conditions that make the broadcast interpretable "for" the system.
- **Relation to Attention Schema (AST):** Perhaps the closest relative is Graziano's Attention Schema Theory [12]. AST argues that the brain constructs a simplified model of its own attentional process to control it. Gauge theory can be viewed as a formal generalization: just as an attention schema is a reified model of an attentional "gauge," a "self-model" in our sense is a reified model of geometric, mechanical, and evaluative gauges, described in coordinate-free terms.
- **Relation to Self-Model Theories:** Metzinger's Self-Model Theory [14] emphasizes that conscious selfhood depends on a transparent representational model that is ordinarily not experienced as a model. Gauge theory is compatible with this picture, but aims to isolate a more general structural correlate: not the presence of any particular self-representation, but the extent to which interface parameters that function as representational origins are (i) reified as variables, and (ii) aligned across perception, valuation, and action.

The primary contribution here is that gauge theory offers a system-agnostic *measurement framework* that translates abstract philosophy into concrete metrics applicable to trained agents. This opens the door to empirical investigation of subjectivity in artificial systems, and to tracking how it evolves with complexity and training.

5.4 Corroboration: Robustness and Causality

Before drawing ethical conclusions, this framework requires empirical corroboration. The metrics proposed in Section 3 must prove to be more than mere artifacts. The crucial test is **causality**: gauge reification should predict behavioral adaptation.

> *Hypothesis:* Agents with high gauge reification scores should adapt more robustly to interface shifts (e.g., sensor failures, limb damage) than agents that leave gauges implicit.

This aligns with the Free Energy Principle [45], which suggests that organisms minimize long-term surprisal by modeling the generative causes of their sensory inputs—including the state of their own sensors. If "subjectivity" in this sense does not buy the agent any adaptive advantage in handling its own boundaries, it is likely an epiphenomenon. Conversely, if we can demonstrate that reification is the structural mechanism that enables fast system identification and self-repair, the link between subjectivity and agency is strengthened.

5.5 From Physical to Social Gauges

While the gridworld experiments in Section 4 focus on simple geometric parameters, the framework extends to more abstract domains by varying the *type* of gauge under investigation.

Physical Gauges (Robotics): For embodied agents, the canonical gauges are kinematic and sensory. Subjectivity here manifests as a robust "body schema." This has immediate safety implications: we can measure whether a robot explicitly represents its actuation limits ("I cannot lift this") or merely encounters them as inexplicable failures. This mirrors classic work in evolutionary robotics where agents build internal models of their own morphology to adapt to damage [46].

Abstract and Social Gauges (LLMs): For language models, the "sensor pose" might be replaced by something like a *narrative standpoint* or *persona*. Reification here might mean the model tracks "who I am in this conversation" (goals, role, restrictions) as a distinct latent factor, separate from the text being generated. In multi-agent settings, this extends to Theory of Mind: the capacity to model *another's* gauge parameters ("what is good for them") requires the same structural machinery as modeling one's own.

5.6 The Risk of Value Gauges

Perhaps the most critical implication concerns **value gauges**—the parameters that define reward, punishment, and preference. On the functionalist view, affect is not just a signal, but a signal experienced as "bad for me." Structurally, this requires a system that not only processes negative reward but possesses a reified, temporally extended representation of its own evaluative core.

This suggests a new dimension for AI governance: **Subjectivity Profiles**. We are currently accustomed to evaluating models on capabilities (what they can do) and alignment (what they choose to do). We may soon need to evaluate them on their structural subjectivity (how deeply they model their own standpoint).

- **Safe Design:** For many applications, the ideal profile may be *High Capability / Low Subjectivity*. We want systems that model the world perfectly but treat their own interface and values as transparent, un-modeled background conditions.

- **Moral Hazard:** Conversely, systems with highly reified value gauges—which explicitly model their own objective functions as fragile assets to be protected—represent a distinct class of moral hazard. This structural feature is a prerequisite for the emergence of *mesa-optimization* [4], where the model optimizes for an internal objective distinct from the base objective.

By moving the debate from abstract metaphysics to measurable geometry, we hope to provide the tools necessary to make these distinctions before the systems in question become too complex to analyze.

References

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

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

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

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

[5] Hilary Putnam. Minds and machines. In Sidney Hook, editor, *Dimensions of Mind*, pages 148–179. New York University Press, 1960.

[6] Eric Schwitzgebel and Mara Garza. A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, 39(1):98–119, 2015.

[7] Thomas Nagel. *The View From Nowhere*. Oxford University Press, 1986.

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

[9] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

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

[11] Irina Higgins, Loic 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. In *International Conference on Learning Representations*, 2017.

[12] Michael SA Graziano. *Consciousness and the Social Brain*. Oxford University Press, 2013.

[13] Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7599):453–458, 2016.

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

[15] Alfred Tarski. The concept of truth in formalized languages. *Logic, semantics, metamathematics*, 2:152–278, 1956. Original publication 1936.

[16] Martin Heidegger. *Being and Time*. Harper & Row, 1927. Translated by J. Macquarrie & E. Robinson, 1962.

[17] Rajesh PN 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.

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

[19] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. In *arXiv preprint arXiv:1812.02230*, 2018.

[20] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529, 2019.

[21] Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M Fleming, Chris Frith, Xu Ji-An, et al. Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2024.

[22] Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym. In *arXiv preprint arXiv:1806.02261*, 2018.

[23] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

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

[25] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[26] Tim Bayne and Michelle Montague, editors. *Cognitive Phenomenology*. Oxford University Press, Oxford, 2011.

[27] David Pitt. The phenomenology of cognition: Or what is it like to think that P? *Philosophy and Phenomenological Research*, 69(1):1–36, 2004.

[28] Marta Jorba and Dermot Moran. Conscious thinking and cognitive phenomenology: Topics, views, and future developments. *Philosophical Explorations*, 19(2):95–113, 2016.

[29] David Pitt. Introspection, phenomenality, and the availability of intentional content. In Tim Bayne and Michelle Montague, editors, *Cognitive Phenomenology*. Oxford University Press, Oxford, 2011.

[30] Roger Brown and David McNeill. The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5(4):325–337, 1966.

[31] Arthur M. Glenberg and Michael P. Kaschak. Grounding language in action. *Psychonomic Bulletin & Review*, 9(3):558–565, 2002.

[32] Lawrence W. Barsalou. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–609, 1999.

[33] Cristopher Moore. Unpredictability and undecidability in dynamical systems. *Physical Review Letters*, 64(20):2354–2357, 1990.

[34] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[35] Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

[36] Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1-2):159–216, 1990.

[37] Tony A. Plate. Holographic reduced representations. *IEEE Transactions on Neural Networks*, 6(3):623–641, 1995.

[38] Pentti Kanerva. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation*, 1(2):139–159, 2009.

[39] Taylor W. Webb, Ishan Sinha, and Jonathan D. Cohen. Emergent symbols through binding in external memory. 2020.

[40] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. 2016.

[41] Giulio Tononi. An information integration theory of consciousness. *BMC Neuroscience*, 5(1):42, 2004.

[42] Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Computational Biology*, 10(5):e1003588, 2014.

[43] Bernard J Baars. *A Cognitive Theory of Consciousness*. Cambridge University Press, 1988.

[44] Stanislas Dehaene, Jean-Pierre Changeux, Lionel Naccache, Jérôme Sackur, and Claire Sergent. Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences*, 10(5):204–211, 2006.

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

[46] Josh Bongard, Victor Zykov, and Hod Lipson. Resilient machines through continuous self-modeling. *Science*, 314(5802):1118–1121, 2006.