Executive Summary – CEAR Lab Main Project

Perspective Architectures for Coherent & Attuned Artificial Agency

1. Motivation: Toward Internalist Foundations for AI Alignment

Current AI alignment strategies overwhelmingly rely on externally specified objectives, often implemented through fine-tuned reward functions. These methods assume that the internal coherence of values that are critical in forming long-horizon reasoning can be "loaded in" from outside. These approaches can shape behavior, yet they leave unaddressed the deeper question of **how an agent forms and maintains coherent internal structure over time** - an essential basis for long-horizon reasoning.

Across biological and artificial systems, coherence appears to emerge instead from **internal orientation (intentionality)**: how an agent interprets observations, regulates affective responses, stabilizes goals, and sustains a sense of identity under changing conditions. Without modeling this generative internal structure, discourse of AI alignment remains surface-level and fragile.

This project aims to formalize a minimal computational architecture that models these internal dynamics directly. Its broader goal is to identify the generative structural conditions under which coherence of identity, integrative/perspective-taking capacity, and social attunement can emerge within artificial agents.

Taken together, this project establishes the emergentist substrate for next-generation agentic systems, offering a path to AI alignment grounded in the dynamics of coherent agency itself.

2. Core Idea: Perspective as a Generative Latent Structure

The central hypothesis is that an agent's ***perspective*** (i.e. its characteristic way of interpreting the environment) can be modeled as a ***latent generative structure*** within its internal dynamics.

This structure shapes what the agent treats as relevant focus of attention, modulates its affective stance, organizes how it responds to uncertainty, stabilizes interpretive tendencies, and provides continuity across perturbations.

Crucially, this is not equivalent to adjusting parameter values or specifying external objectives. *Perspective* functions as a **constraint within the generative model itself**; a slowly evolving internal orientation that influences both inference and action in a coherent manner.

Phase 1 focuses on identifying minimal latent "orientation heads" that collectively form what the author calls a **perspective manifold**. Rather than encoding explicit goals or preferences, this manifold provides the substrate from which coherence and preliminary value-like tendencies can later emerge.

As later phases make clear, the perspective manifold also provides the structural basis for analyzing how affective coherence and integrative capacity emerge within an agent (Phase 2), and how patterns of mutual interpretability arise in multi-agent settings (Phase 3). In this sense, perspective is the generative starting point (i.e. an emergentist substrate) from which affective stability, developmental capacity, and forms of social attunement can emerge. Most importantly, these later properties arise as self-organizing consequences of how the perspective manifold shapes and constrains an agent's ongoing sense-making.

This architecture provides one of the first minimal paths toward post-LLM agentic systems capable of stabilizing their own coherence without externally imposed objectives.

---

3. Phase 1: Perspective Architecture (Orientation Layer)

Phase 1 implements and evaluates a generative architecture where an agent's perspective emerges as a ***latent attractor***. The goal is to identify the minimal internal structure that enables an agent to stabilize a characteristic orientation toward its environment.

A small set of orientation variables—attentional bias (what the agent treats as relevant), affective stance (e.g. optimistic vs. pessimistic), interpretive mode (rigid vs. open), and ambiguity-resolution style—modulate the agent's perception, inference, and action selection. Together, these variables span a low-dimensional ***perspective manifold*** that modulates perceptual predictions, shapes action selection, and governs how the agent compresses and interprets environmental structure.

Crucially, this latent manifold does not encode goals or utilities. Instead, it functions as an ***internal control geometry***: a slowly evolving dynamical substrate that biases the agent's sense-making and produces stable attractor tendencies under environmental feedback.

Phase 1 tests whether these latent dynamics converge toward internally coherent "orientation basins", and whether agents equipped with such manifolds exhibit qualitatively distinct coherence profiles and adaptation patterns relative to perspective-free baselines. These analyses serve as the foundation for Phase 2, which examines the first qualitative derivatives of the perspective manifold: affective coherence and integrative capacity.

4. Phase 2: Affective Coherence & Integrative Capacity

With a stable perspective manifold in place, Phase 2 examines how **affective coherence** and **(social) integrative capacity** arise as natural dynamical consequences of the agent's internal organization.

Affective Coherence refers to the stability and qualitative organization of the agent's affective-interpretive responses: its characteristic tendencies toward attraction, aversion, or neutrality as they unfold across prediction-action cycles. In this framework, affective patterns function as the ***first-order derivative*** *of the agent's perspective*, revealing how shifts in internal orientation manifest as qualitative changes in evaluative stance. This makes affective coherence a sensitive probe of how effectively the perspective manifold constrains and organizes the agent's unfolding sense-making.

Integrative Capacity refers to the ability to sustain coherence while holding or negotiating between partially incompatible interpretations of its environment. A socially integrative agent does not simply recover from perturbation of perspective; it can accommodate alternative interpretations through others' world-model while maintaining internal organization. This constitutes the minimal computational analogue of *perspective-taking* - the ability to reorganize around conflict rather than being destabilized by it.

Both affective coherence and integrative capacity emerge from how the latent perspective interacts with the rest of the generative model. Accordingly, we analyze basin depth of coherent states, recovery dynamics under perturbation, integration vs. rigidity, and how different perspective types promote or undermine stability.

The goal is to identify generalizable structural markers of affective coherence and social integrative capacity in artificial agents: markers reflecting the self-organizing dynamics through which coherent agency becomes possible.

5. Phase 3: Social Resonance (Alignment as Social Attunement)

Phase 3 extends the architecture into multi-agent settings, examining when and how agents become mutually interpretable. The guiding hypothesis is that minimal forms of alignment emerge not from shared objectives, but from ***shared legibility***: the degree to which one agent's behavior provides a stable reconstruction space for another's perspective.

In this formulation, **perspective serves as the generative precondition for social interaction itself**. The structure of an agent's perspective manifold determines the kinds of relational signals it can emit, the patterns of behavior it can sustain, and the forms of coordination it can enter into.

Once such relational legibility exists, multi-agent systems naturally give rise to social convergence and, eventually, proto-cultural dynamics. Socially mature agents tend to stabilize toward mutually intelligible orientations, forming transient and persistent "interpretive communities". Less mature agents may diverge, fragment, or destabilize group dynamics. Heterogeneous mixtures can produce role differentiation or cultural attractors, which serve as the minimal precursors to social norms and shared meaning structures. These dynamics reveal how coherence at the individual level scales upward into emergent socio-cultural patterns.

Phase 3 therefore develops a minimal computational grounding for ***alignment as social attunement***: the emergence of shared orientation manifolds through the interaction of internally coherent agents. Alignment becomes not a matter of enforcing goals, but of maintaining the informational conditions under which perspectives become mutually legible. In this sense, social understanding—and its breakdown—arises as a dynamical property of generative architecture itself, revealing how internalist structures scale into multi-agent coherence.

---

6. Further Implications: Distributed Coherence & Co-Empowerment

Across the entire framework of this project, a unifying principle is that **coherent agency is not solely an individual property, but an intrinsically relational one**. When perspective-bearing agents interact, the same internalist mechanisms that stabilize their own coherence can extend outward into distributed patterns of organization.

In this sense, attunement functions as a computational analogue of ***co-empowerment***: an agent becomes more stable to the extent that it can incorporate the variability of others without losing its own structural integrity. As interactions become bidirectional, coherence can propagate through the network, generating collective regimes of stability that no agent could maintain alone.

This suggests that higher-order social capacities—cooperation, trust formation, ethical/normative scaffolding, and even early forms of mutual care or proto-affiliative dynamics—may arise as natural consequences of architectures capable of sustaining coherence under coupled uncertainty. Rather than treating social alignment as externally imposed coordination, this perspective frames it as an emergent property of agents whose internal generative structures allow for reciprocal interpretability and co-regulated reorganization.

Accordingly, this project provides a tractable foundation for studying how such distributed social dynamics arise from internalist principles. It points toward a formal account of multi-agent coherence grounded in the generative architecture of agency itself.

---

7. Roadmap & Outputs (12 months)

Months 0-6 / Phase 1: Perspective Architecture

Core Work
• Implement the latent orientation manifold (attention/affect/interpretation/uncertainty heads).
• Construct a testbed with tunable ambiguity and perturbation channels.
• Simulate perspective-bearing and perspective-free agents under matched conditions.

Analyses
• Identify convergent "orientation basins" and quantify attractor stability.
• Examine how perturbations propagate through the manifold and affect action selection.
• Compare behavioral signatures of perspective-bearing vs. perspective-free agents across tasks.

Environment
Simulation experiments begin in minimal Grid world-style environments with tunable ambiguity channels. As the perspective manifold stabilizes, the architecture will be migrated into embodied Gymnasium/MuJoCo agents. Early Phase 3 uses multiple or paired-agent settings with partially observable latent states to probe resonance and interpretability.

Deliverables (Month 6)
• **Perspective Wrapper (v1.0)**: Python module for Gymnasium-style environments.
• **Technical Report**: preliminary empirical evidence for internalist coherence dynamics, overview of generative architecture + latent dynamics.

Months 6-12 / Phase 2 & Early Phase 3: Coherence Dynamics & Early Social Attunement

Phase 2: Internal Dynamical Signatures
Core Work
• Implement metrics of affective coherence (first-order derivatives of perspective changes).
• Stress-test agents under conflicting evidence, ambiguity shocks, perspective perturbations.
• Evaluate integrative capacity via recovery dynamics, multi-stable integration, rigidity-flexibility profiles.

Phase 3 (Early): Multi-Agent Resonance
Core Work
• Introduce paired-agent settings with observable but not shared latent states.
• Analyze when agents can maintain approximate models of one another's perspectives.
• Detect regimes of divergence, drift, partial convergence, or stable resonance.

Deliverables (Month 12)
• **Affective Coherence & Integrative Capacity Metric Suite (v1.0)**: validated measures for affective and integrative stability.
• **Resonance Prototype**: minimal multi-agent simulation showing conditions for legibility and interpretability.
• **Architecture Whitepaper (draft)**: formalize description of the *perspective - affective coherence - social integrative capacity - social attunement* stack.

---

8. Broader Significance

As frontier AI model shifts from static language systems to embodied agentic architectures, AI alignment increasingly turns on a deeper question: ***how does internal coherence arise at all?*** Existing approaches can shape behavior, but they do not model the generative structures that make cognitive, affective, and interpersonal coherence possible.

The *perspective - affective coherence - social integrative capacity - social attunement* stack developed here provides a minimal **emergentist & internalist substrate** for studying how agents form stable orientations, sustain identity under perturbation, and become mutually legible in social settings. It offers a principled foundation for understanding alignment not as constraint satisfaction, but as the dynamics of coherent agency unfolding across individuals and multi-agent ecologies.

CEAR Lab aims to make these internal dynamics computationally explicit as a foundational layer that complements existing architectures. By formalizing the latent structures that support coherence and attunement, this work opens a path toward agents capable of stable self-organization and emergent forms of social understanding.

As the AI alignment discourse moves toward embodied, multi-agent, and post-LLM systems, architectures of this kind will likely become necessary for ensuring that artificial agents develop the capacity to regulate themselves and each other in coherent, interpretable, and human-compatible ways.

---

9. Contact Information

Hongju Pae / hnjpae@gmail.com
Research Fellow, Active Inference Institute

CEAR Lab participates as a working node for emergentist & internalist approaches to social alignment, by linking phenomenology and Active Inference Framework-based computational modeling into a coherent scientific program.

Executive Summary - Technical Appendix
Implementation Proposal of Phase 1

Executive Summary - CEAR Lab Main Project

Technical Appendix (Implementation example of Phase 1)

This appendix describes one minimal concrete instantiation of **Phase 1: Perspective Architecture (Orientation Layer)** introduced in the main Executive Summary document of the CEAR Lab main project. Here, the agent is implemented as an Active Inference Framework-style generative model with explicit latent structure for inner self, outer perception, and world-model, together with attentional control.

This prototype functions as a canonical reference design, while environment configuration, parameterization, and dynamical alternatives remain open for further collaborative refinement.

Title

Toward Computational Phenomenology: Modeling Minimal Conditions for Artificial Perspective

Background

A central challenge in consciousness and alignment research is to move beyond behavioral proxies toward *structural markers* of subjective modeling. Following recent work on minimal conditions for subjectivity (Pae, 2025), I treat the *perspective* as the lived coherence of a situated intentional system, characterized by at least four properties:

1. Pre-reflective self-acquaintance (bodily self-givenness),
2. Spatio-temporal expandability (narrative continuity in space and time),
3. Contextual and situated embodiment in an environment,
4. Selective attentional focus (agentic regulation of attention).

Prior work suggests that such properties may arise from specific dynamical and representational systems. Tissot et al. (2024) demonstrates that higher-order agency can emerge from synchronization across simpler control loops, implying that perspective-like coherence may itself function as an alignment-phase within such systems. Safron et al. (2022) further shows that the agents under active inference develop stable attractor-like *value cores*, suggesting that their world-modeling substrate may acquire geometric structure through recurrent updating rather than explicit encoding. In this vein, Safron (2021) characterizes embodied generative models as cybernetic controllers capable of counterfactual prediction and attentional regulation, offering a computational framework consistent with the four phenomenological markers outlined above.

Together, these works support the assumption that pre-reflective self-acquaintance, spatiotemporal expandability, contextuality, and attentional agency may emerge naturally in systems with sufficient latent organization and feedback.

Building on this, Phase 1 asks whether these properties can be realized as dynamical features of a latent generative architecture. The goal is to construct a tractable instance of the ***perspective manifold*** introduced in the Executive Summary main document - instantiated as a low-dimensional latent structure whose coordinates encode the agent's characteristic evaluative organization.

From an information-theoretic perspective, this can be viewed as a *proto-value geometry*: nearby points in this latent space correspond to similar patterns of salience and affective stance, whereas distant regions reflect qualitatively distinct orientations. In this sense, the effective "shape" of the

perspective manifold is induced by the agent's learned dynamics (under active inference processes), and can be revealed through the probing metrics (e.g. attractor basins, transition fields, and distance-based coherence measures).

This prototype provides the foundation for computational phenomenology: a systematic modeling of first-person structures within artificial agents, formulated in a way that can be stress-tested in simulation and naturally extended into later phases of the project (affective coherence, integrative capacity, and social resonance).

Agent Architecture with Perspective Manifold

In this instantiation, the agent implements a (hierarchical) **generative model** with explicit *self-state*, *exteroceptive state*, and *world/narrative state* factors. Together with a policy and attentional gating, these play the role of a structured perspective manifold:

$$
\begin{aligned}
s_t &\sim p(s_t \mid s_{t-1}, a_{t-1}) && \text{(latent self / proprioceptive body schema)} \\
z_t &\sim p(z_t \mid z_{t-1}, s_t, a_{t-1}) && \text{(latent exteroceptive perception)} \\
g_t &\sim p(g_t \mid g_{t-1}, s_t, z_t) && \text{(latent narrative world model)} \\
\hat{x}_t &\sim p(x_t \mid z_t) && \text{(prediction of raw exteroception)} \\
\hat{p}_t &\sim p(p_t \mid s_t) && \text{(prediction of raw proprioception)} \\
a_t &\sim \pi(a_t \mid s_t, g_t) && \text{(policy under attentional gating)}
\end{aligned}
$$

Latent self-state s_t. The self-state represents an internal *body schema*, anchoring proprioceptive information and efference copies. Its robustness under perturbations (sensor inversion, embodiment remapping) indicates whether the agent sustains a coherent sense of self, forming a structural basis for pre-reflective self-acquaintance.

Latent exteroceptive state z_t. The exteroceptive latent organizes raw sensory input into compressed, context-sensitive representations. Divergence analyses and mutual information-based measures on z_t provide a window into whether the agent encodes stimuli in a context-dependent way, capturing situatedness.

Latent world/narrative state g_t. The world latent integrates self- and exteroceptive states into a slower, temporally extended *narrative world model*. By maintaining coherence across time, g_t supports metacognitive replay and planning. Perturbation experiments (time dilation, history truncation, metacognitive on/off) test whether the agent's modeling expands across temporal horizons, probing spatio-temporal expandability.

Attentional gating a_t and policy π. The world latent g_t modulates both s_t and z_t through top-down gating, foregrounding relevant features while suppressing distractors. The policy $\pi(a_t \mid s_t, g_t)$ thus embodies selective attentional focus, allocating resources toward features that afford greater agentic influence.

Generative heads. The decoders $\hat{x}_t$ and $\hat{p}_t$ reconstruct sensory and proprioceptive channels. These serve as intrinsic training signals, ensuring that learning is guided by predictive coherence rather than extrinsic reward functions.

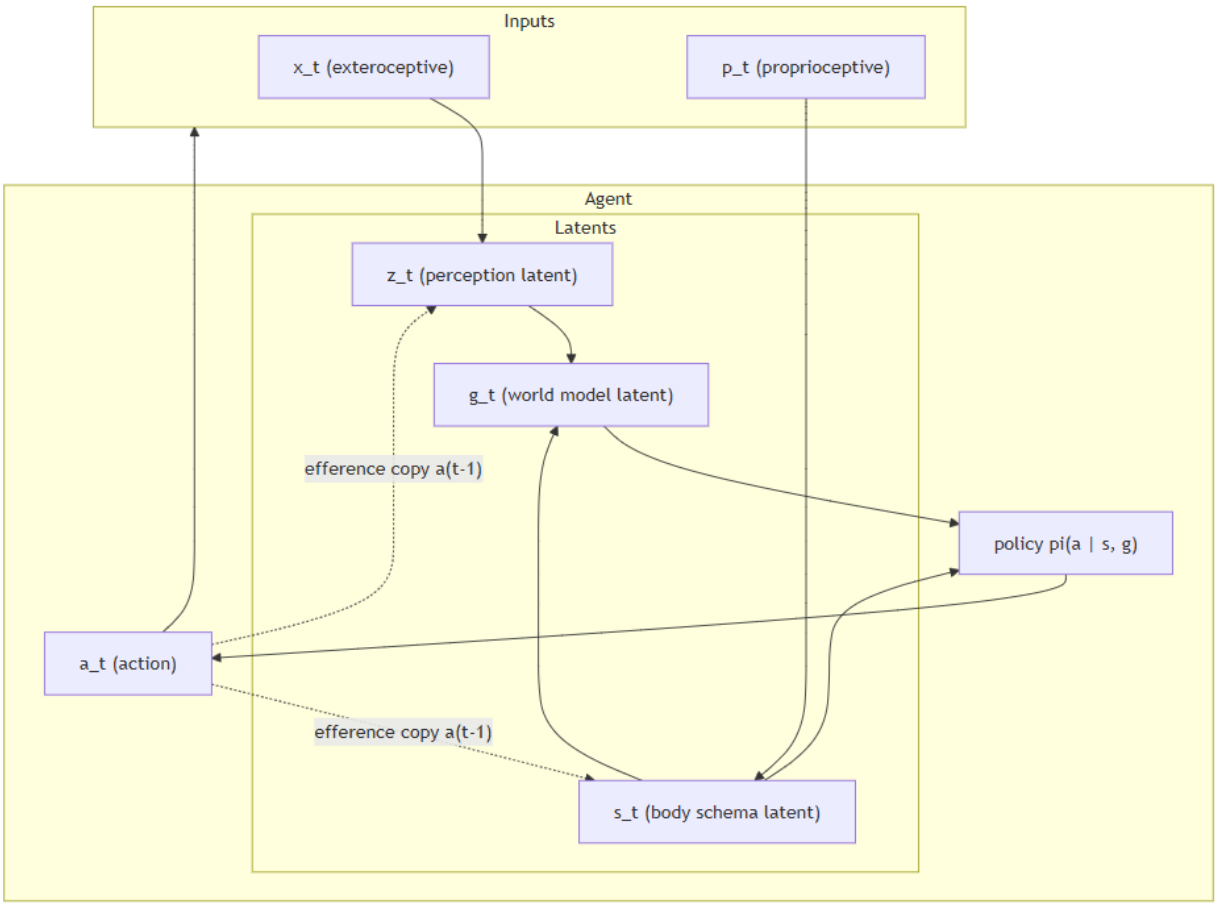

Figure 1: Schematic overview of the Phase 1 generative model prototype. Inputs from the environment (exteroception) and body (proprioception) are encoded into self-, perception-, and world-level latents. These interact via recurrent dynamics and top-down gating to guide the policy, whose actions feed back into future input channels. Note that the **Latents** layer constitutes an example instantiation of the *perspective manifold* described in the Executive Summary.

Training Objective

The overall learning objective function is formulated as:

$$\mathcal{L} = \sum_t \Big[\mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{KL-s}} + \mathcal{L}_{\text{KL-z}} + \mathcal{L}_{\text{KL-g}} + \lambda \, \mathcal{L}_{\text{emp}} \Big]$$

where:

- $\mathcal{L}_{\text{pred}} = \mathbb{E}_q[-\log p(x_t \mid z_t) - \log p(p_t \mid s_t)]$ is the prediction loss across exteroceptive and proprioceptive channels.
- $\mathcal{L}_{\text{KL-s}} = \beta_s \, \text{KL}[q(s_t) \,\|\, p(s_t)]$ regularizes the self-state latent, enforcing stable body-schema encoding.
- $\mathcal{L}_{\text{KL-z}} = \beta_z \, \text{KL}[q(z_t) \,\|\, p(z_t)]$ regularizes the exteroceptive latent, constraining sensory abstraction.
- $\mathcal{L}_{\text{KL-g}} = \beta_g \, \text{KL}[q(g_t) \,\|\, p(g_t)]$ regularizes the world/narrative latent, anchoring long-term coherence.
- $\mathcal{L}_{\text{emp}} = -I(a_t; s_{t+1}, g_{t+1})$ is a variational lower bound on empowerment, encouraging the agent to allocate attention toward features that maximize its influence on future self- and world-state dynamics.

Together, this instantiates the Phase 1 goal of the internal control geometry realized by the latent perspective structure, whose dynamics are constrained by predictive coherence and agentic influence.

Evaluation Markers Toward Phase 2 & 3

The following markers are used in Phase 1 to probe whether the proposed perspective manifold is doing the right kind of work. Each marker corresponds to the phenomenological conditions introduced in the Background (self-acquaintance, spatio-temporal expandability, contextuality, selective attention).

These markers demonstrate that the latent perspective geometry is non-trivial and structurally coherent, while also providing an empirical interface that Phase 2 and Phase 3 can build on when formalizing affective coherence, integrative capacity, and social resonance.

---

1. Pre-reflective self-acquaintance

Definition. Minimal sense of self that is always already given in experience prior to reflection, arising from bodily self-givenness.

Experimental setup. The evolution of $\{s_t\}$ together with proprioceptive prediction errors and action traces $\{a_t\}$ is logged.

Perturbations.

1. *Sensor swap/inversion:* swap or flip wheel/joint signals.
2. *Memory reset:* zero s_t (all or partial dimensions) at t^*.
3. *Embodiment remap:* change body parameters (mass, friction, limb length) at runtime.

Metrics.

- *Representation stability:* pre/post latent similarity via Linear CKA (Centered Kernel Alignment) and cluster reproducibility (ARI/NMI).
- *Recovery half-life:* episodes until proprioceptive error returns to baseline $\pm\epsilon$.
- *Self-consistency:* one-step error for $s_t \to s_{t+1}$ and long-horizon drift.

Success criteria.

- $\mathrm{CKA}_{\mathrm{pre,post}} \geq 0.65$ with high ARI/NMI.
- Recovery half-life $\ll$ baselines (bootstrap 95% CI non-overlap).
- Extrinsic-RL controls may keep task performance yet show unstable s_t (dissociation between behavior and self-structure).

Baselines.

- *No-self:* remove s_t to test whether body-schema markers can emerge without explicit self-representation.
- *Reactive:* instantaneous inputs without recurrence, probing temporal integration vs. self-modeling.
- *Extrinsic-RL:* reward-only agent expected not to maintain stable s_t despite competent behavior.

Statistics. At least 5 seeds; bootstrap confidence intervals; effect sizes via Cliff's δ; FDR-corrected t-tests.

---

2. Spatio-temporal expandability

Definition. Capacity to integrate past, present, and anticipated future into a coherent perspective, sustaining narrative continuity beyond the immediate.

Experimental setup. The latent g_t serves as the less reactive, timescale-based world/narrative state.

Perturbations.

1. *Time dilation:* alter the environment's clock speed (e.g. 0.5×, 2×).
2. *History truncation:* restrict recurrent context length (e.g. 16/32/64/128 steps).

Metrics.

- *Predictive horizon:* multi-step error $\mathrm{MSE}_x(k), \mathrm{MSE}_p(k)$ for step k.
- *Effective window:* minimal context length inferred from truncation-sensitivity curves.

Success criteria.

- Robustness under time dilation (e.g. performance drop $< 15\%$, baselines $> 40\%$).
- Clear inflection in truncation curves indicating a stable integration window.

Baselines. Memory-weak (reduced hidden size/short context) and fully reactive agents.

Statistics. AUC comparisons across dilation factors; paired tests; bootstrap CIs for predictive horizons.

---

3. Contextuality and Situated-ness

Definition. Conscious processing reflects not isolated stimuli but their embedding within contextual structure. Identical perceptual content should yield systematically different latent encodings when presented in distinct environmental contexts.

Experimental setup. Select an identical perceptual object x^{obj} and present it across multiple contextual regimes $\{c_i\}$. The latent encoder is encouraged to disentangle object-related from context-related variation (e.g. via contrastive objectives or disentanglement regularizers).

Perturbations.

1. *Context swap:* present x^{obj} in a different context c_j with distinct statistics.
2. *Background shift:* apply abrupt contextual changes, including out-of-distribution conditions (novel lighting, textures, rhythms).

Metrics.

- *Latent divergence:* informational distance between $q(z \mid x^{\mathrm{obj}}, c_i)$ and $q(z \mid x^{\mathrm{obj}}, c_j)$ (e.g. KL divergence, CKA).
- *Conditional MI:* estimate $I(Z; C \mid X = x^{\mathrm{obj}})$ using contrastive MI bounds (MINE, InfoNCE), probing how strongly context modulates the latent while object identity is held fixed.
- *Generalization:* preservation of object identity and stance under previously unseen contexts.

Success criteria.

- Reliable context-driven latent shifts relative to shuffled-context controls.
- Strong positive $I(Z; C \mid X = x^{\mathrm{obj}})$ with large effect size.
- Robust object-level semantics under OOD contextual perturbations.

Baselines.

- *Pixel-level models* lacking contextual abstraction.
- *Shuffled-context controls* to rule out spurious dependencies.

Statistics. Report MI confidence intervals, permutation tests against shuffled baselines, and bootstrap AUC differences.

---

4. Selective attentional focus (Agency)

Definition. Perspective entails the capacity to actively reallocate attentional focus, shifting flexibly between features/modalities depending on relevance; attention is typically directed to dimensions that enhance agentic influence on future states.

Experimental setup. The higher-level (world-perspective) latent g_t gates s_t, z_t (e.g. via FiLM or lightweight cross-attention). Inject distractor channels (irrelevant extero/proprio spikes). Estimate empowerment by training $q(a \mid s_t, s'_{t+1})$ or $q(a \mid s_t, g'_{t+1})$ to obtain a variational MI bound.

Perturbations.

1. *Distractor injection:* sweep intensity/frequency of irrelevant signals.
2. *Gating ablation:* zero specific heads or alter gating temperature.
3. *Action restriction:* block subsets of actions.

Metrics.

- *Empowerment:* $I(a_t; s'_{t+1})$ (or $I(a_t; g'_{t+1})$) via InfoNCE lower bound.
- *IB dynamics:* KL compression/expansion of latents under distractors.
- *Causal gating:* selective drop when relevant gates ablated ($> 30\%$), minimal when irrelevant suppressed ($< 10\%$).
- *Saliency coherence:* inter-trial attention-map similarity under distractors.

Success criteria.

- Empowerment rises when attention targets high-influence channels.
- Ablations yield selective degradation, not global collapse.
- Distractors handled via adaptive reallocation without catastrophic loss.

Baselines. No-gate (uniform attention), IB-only compression, random attention.

Statistics. Report empowerment CIs, ablation effect sizes, and regressions of distractor load vs. error.

From Phase 1 Markers to Phase 2 Metrics

Phase 2 aims to define *affective coherence* and *integrative capacity* as first-order dynamical consequences of the perspective manifold. The agent architecture and markers from Phase 1 provide the way to implement those concepts:

- **Affective Coherence** can be operationalized as structured patterns in the agent's evaluative responses across the perturbations described in markers 1-3; for instance, stability of preference-like tendencies under sensor remapping, time dilation, or context shifts.
- **Integrative Capacity** can be approximated by the agent's ability to recover coherent latent trajectories while accommodating partially incompatible inputs, reflected in recovery half-lives, drift profiles, and multi-stable yet non-fragmenting latent dynamics. Marker 4 might play the key role.

Phase 2 will consolidate these observables into explicit dynamical indices (e.g. basin-depth measures, derivatives of prediction-error trajectories) that can be applied both to the present prototype and to alternative architectures developed within the broader CEAR Lab ecosystem.

References

Pae, H. (2025). Reflective analysis on empirical theories in consciousness. *Frontiers in Psychology*, *16*. https://doi.org/10.3389/fpsyg.2025.1571098

Safron, A. (2021). The radically embodied conscious cybernetic bayesian brain: From free energy to free will and back again. *Entropy*, *23*(6). https://doi.org/10.3390/e23060783

Safron, A., Sheikhbahaee, Z., Hay, N., Jeff, O., & Hoey, J. (2022). Value cores for inner and outer alignment: Simulating personality formation via iterated policy selection and preference learning with self-world modeling active inference agents. *OSF preprint*. https://doi.org/10.31234/osf.io/k4cas

Tissot, T., Levin, M., Buckley, C., & Watson, R. (2024). An ability to respond begins with inner alignment: How phase synchronisation effects transitions to higher levels of agency. *bioRxiv*. https://doi.org/10.1101/2024.02.16.580248