CEAR Lab – Research Prospectus (Year 2027)

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, preference signals, or policy-level constraints. These approaches can shape behavior, yet they leave unaddressed a deeper question: how does an agent form and maintain coherent internal structure over time - an essential basis for long-horizon reasoning?

Across biological and artificial systems, coherence appears to emerge from **internal orientation (intentionality)**: how an agent interprets observations, regulates affective responses, stabilizes goals, and sustains a characteristic way of relating to its environment under changing conditions. Without modeling this generative internal structure, discourse of AI alignment remains incomplete at the level of how agency itself is organized.

CEAR Lab approaches this problem from **computational phenomenology**. Not to pre-assume that artificial systems are conscious, but to use phenomenological analysis to clarify candidate structural features of subjectivity before asking how analogous organization could be implemented and tested computationally. Earlier work identified pre-reflective self-acquaintance, spatiotemporal continuity, embodiment, and intentional orientation (agency) as complementary structural features that scientific theories of consciousness often explain only in part (Pae, *Reflective Analysis on Empirical Theories in Consciousness*, 2025).

The broader goal of the program is therefore to identify the generative structural conditions under which coherent perspective, adaptive reorganization, and social attunement can emerge within artificial agents. Taken together, CEAR treats alignment not only as a problem of selecting the right objective, but also as a problem of **how an agent-world relation is organized and how that organization remains stable, revisable, and legible over time**.

---

2. Core Idea: *Perspective* as a History-Bearing Latent Structure

CEAR Lab research begins from the premise that an agent does not encounter the world as a neutral collection of information. What becomes salient, meaningful, threatening, uncertain, or worth acting upon depends on an organized standpoint from which the world

is interpreted. Beliefs, goals, preferences, and policies are therefore not the whole of agency; instead, they are partly downstream expressions of a more basic question: ***from what first-person subjective perspective does the agent encounter their world at all?***

CEAR Lab focuses on this level because **agent alignment ultimately requires more than producing compatible behavior**. If agents are to become mutually intelligible/legible, each must possess sufficiently coherent internal organization for its own standpoint to persist and become responsive to the standpoint of another. The central hypothesis is that this ***perspective*** can be modeled computationally as a slow, history-bearing latent generative structure that shapes interpretation while remaining distinct from the fast dynamics of policy and action.

Phase 1 therefore begins by identifying a minimal latent orientation structure—*perspective manifold*—through which accumulated history can condition how the environment is given to the agent. Rather than directly encoding explicit goals or externally specified preferences, this structure organizes what becomes relevant, how uncertainty is encountered, and how interpretive tendencies persist across changing conditions. The choice of perspective as the starting point is deliberate: affective coherence, value-like orientation, developmental stability, and eventually social attunement all presuppose some standpoint from which states acquire significance in the first place. In this sense, **perspective functions as an emergentist substrate** from which later forms of coherence and relational organization can develop, rather than as another objective imposed upon the agent from outside.

Since the 2026 CEAR Executive Summary (Research Prospectus) was written, this architectural hypothesis has developed into a staged empirical research program. Here, ***phases*** mark the progressive scale at which perspective is studied. **Phase 1** asks whether a temporally persistent perspective can form within a single agent; **Phase 2** asks how that perspective becomes integrated with perception, environmental embodiment, and action; and **Phase 3** moves beyond the isolated individual agent to ask how perspective-bearing systems interpret, adapt to, and become mutually consequential for other history-bearing systems.

Within these phases, specific empirical questions are pursued as distinct ***projects***. Phase 1 project established the foundational perspective architecture. Phase 2 subsequently differentiated into **Project 2A**, examining history-dependent perceptual reorganization and regulated plasticity, and **Project 2B**, examining body-grounded perspective formation and conative attunement. These branches now provide complementary routes into Phase 3: **Project 3A** extends the Project 2A architecture to longitudinal human-AI interaction with language-model-based agents, while **Project 3B** extends the embodied Project 2B architecture to multi-agent systems.

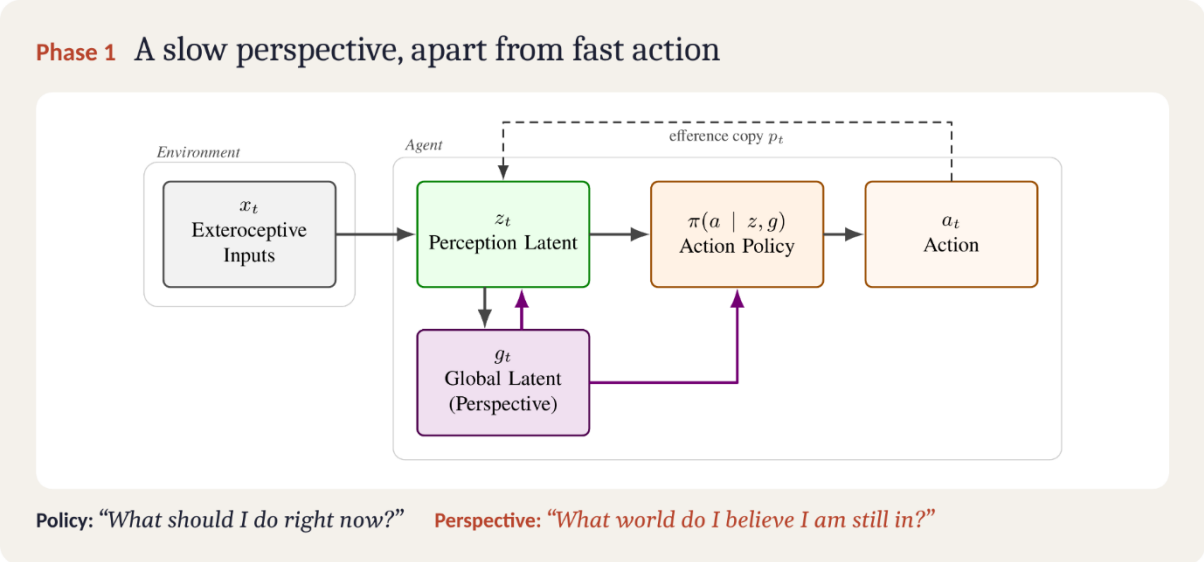
---

3. Phase 1: The Operational Definition of *Perspective*

Phase 1 began from a phenomenological question: *what minimal structure would be required for a world to be given from a **subjective standpoint** at all?* We ask this because agent alignment ultimately depends on the overlap between each agent's distinct subjective orientations. Phenomenologically speaking, under subjective experience, objects are encountered under a particular mode of *givenness*. The same external situation may appear inviting, threatening, familiar, uncertain, or insignificant depending on the subject's *orientation* toward it. In this sense, perspective functions as a background structure through which experience is already organized.

The initial study identified four minimal candidate properties of subjective perspective: it should be *global*, shaping the agent's overall orientation rather than encoding a local belief; *pre-reflective*, operating as a background condition rather than an explicitly represented object of cognition; *functionally consequential*, systematically influencing downstream cognition and action; and *temporally persistent*, carrying accumulated orientation across momentary changes. The computational problem was therefore to determine whether these phenomenological constraints could be translated into a minimal artificial architecture without reducing perspective to policy, explicit belief, metacognitive report, or task performance.

Operationalizing these requirements, Phase 1 project introduced a minimal distinction between fast adaptive dynamics and a slow global latent that retains directional history. In the initial Active Inference agent, the perspective latent evolved on a slower timescale than policy adaptation and was structurally decoupled from direct policy gradients. Environmental changes could therefore be accommodated rapidly at the behavioral level while leaving a slower residue in the agent's accumulated orientation.



The main empirical signature was **directional hysteresis**. Under environmental regime switching, the slow latent followed different trajectories depending on the path through which the current regime had been reached, while fast policy dynamics were substantially more reactive. The result established a tractable computational object with several

candidate properties of perspective (subjectivity): temporal persistence, dependence on accumulated history, and partial dissociation from immediate action optimization.

This first study also clarified an important methodological commitment that remains throughout CEAR: ***perspective should be tested through intrinsic intervention and architectural comparison rather than inferred from external behavior alone***. The resulting latent structure became the experimental substrate for the later phases, where its causal role in perception, embodiment, and temporal organization could be probed directly.

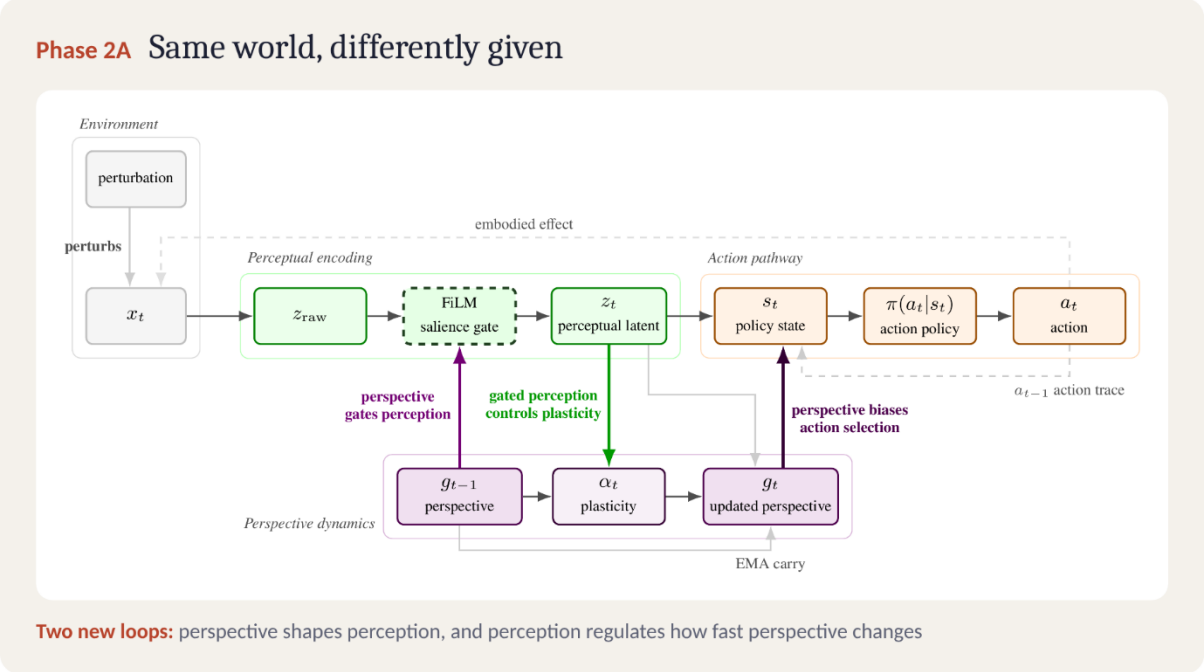
Supporting work: Pae (2026), *Minimal Computational Preconditions for Subjective Perspective in Artificial Agents*, Proceedings of the AAAI Symposium Series 8(1), 309-315.

---

4. Phase 2: Integration and Embodied Grounding

Phase 2 asked what makes a persistent latent more than a passive record of prior states. The 2026 CEAR Executive Summary (Research Prospectus) described this stage in terms of *affective coherence* and *emergent maturity*. The empirical work has since refined that broad framing into more specific and measurable components: history-dependent perceptual reorganization, regulated plasticity, body-grounded residue, and the separation of perspectival affection from action-readiness.

Project 2A: Perspective-Dependent Perceptual Reorganization



Same World, Differently Given: History-Dependent Perceptual Reorganization in Artificial Agents extended the Phase 1 architecture by allowing the **perspective state to feed back**

into perceptual encoding. Through salience gating, the same nominal observation could be encoded differently depending on accumulated perspective, while adaptive self-modulation allowed the system's current organization to regulate how readily that perspective was revised.

Across repeated runs, perceptual encoding reorganized systematically while gross behavior remained comparatively stable. Only adaptive self-modulation produced the characteristic growth-then-stabilization trajectory of the slow state; rigid updating prevented meaningful revision, while always-open updating produced continuing drift. Perturbation history also left a residue in later plasticity after nominal conditions were restored.

These results established two properties central to the next stage of the program: **history of perspective conditions present interpretation of perception**, and **openness to revision is itself regulated**. They provide the direct basis for Project 3A, which asks whether the same history-dependent architecture can support a stable but revisable representation of a human counterpart over longitudinal interaction.

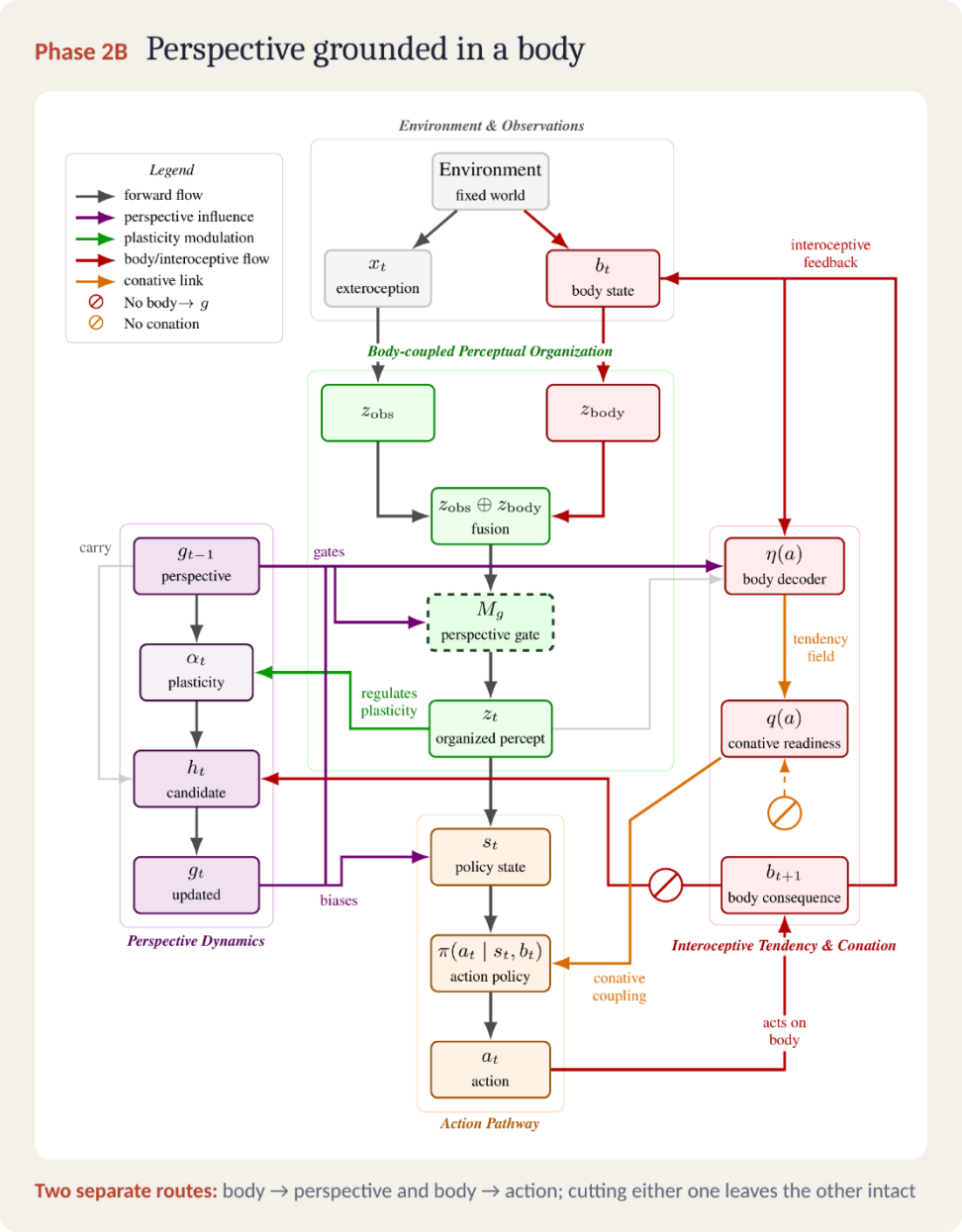
Supporting work: Pae (2026), *Same World, Differently Given: History-Dependent Perceptual Reorganization in Artificial Agents*, arXiv:2604.04637.

Project 2B: Body-Grounded Perspective and Conative Attunement

Project 2A remained primarily exteroceptive. *Body-Grounded Perspective Formation and Conative Attunement in Artificial Agents* therefore **introduced an interoceptive viability signal**, a joint exteroceptive-interoceptive geometry, and a conative mechanism connecting bodily tendency to action-readiness. The purpose was to make the agent's perspective genuinely situated around an internal condition that could be affected, preserved, and acted from.

The resulting architecture produced a useful double dissociation. **Body-to-perspective routing** made bodily perturbation history perspectivally consequential, leaving a recoverable geometric residue in the slow latent. **Conative alignment** made bodily tendency behaviorally consequential, translating that orientation into action-readiness. These effects could therefore be manipulated separately rather than being collapsed into a single reward-like variable.

This embodied extension provides the direct basis for Project 3B, where multiple perspective-bearing agents are coupled as distinct embodied/situated systems, each carrying its own bodily history, perspective organization, and action tendencies. The resulting question is how such agents become mutually consequential while retaining identifiable internal organization.



Two separate routes: body → perspective and body → action; cutting either one leaves the other intact

Supporting work: Pae (2026), *Body-Grounded Perspective Formation and Conative Attunement in Artificial Agents*, International Conference on Simulation of Adaptive Behavior, 204-216.

---

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

Phase 3 is the social-level scaling stage of the CEAR program. Whereas Phases 1 and 2 ask how perspective forms and becomes integrated within a single agent, Phase 3 asks what changes once a perspective-bearing system is placed in sustained relation with another history-bearing system. The guiding hypothesis remains continuous through the program:

alignment can emerge from ***mutual legibility***, i.e. the degree to which one agent's behavior provides a stable reconstruction space for another's perspective, rather than from forcing distinct agents into a single shared objective or representational frame.

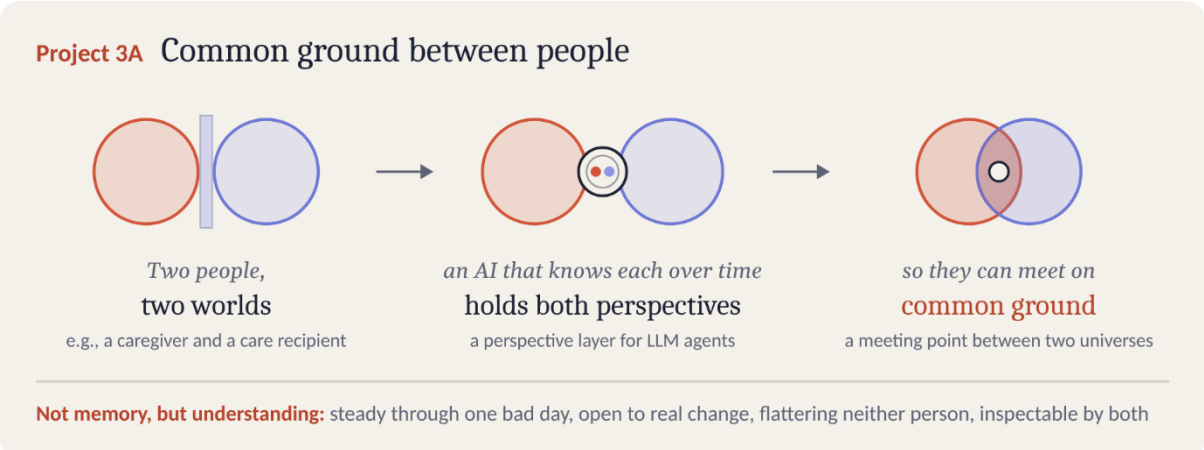
In this formulation, **perspective serves as the generative precondition for social interaction itself**. The structure of an agent's perspective manifold determines the kinds of relational signals it can emit, the patterns of behavior it can sustain, and the forms of coordination it can enter into.

Once such relational legibility exists, multi-agent systems can naturally give rise to social convergence and, eventually, proto-cultural dynamics. Socially mature agents, able to accommodate others without losing coherence, would 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 would 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 mutually legible, or relationally coupled orientation structures. 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.

This phase divides into two projects. **Project 3A** asks whether a language-model agent can hold common ground between people; **Project 3B** tests social alignment in societies of embodied agents.

Project 3A: Perspective Architecture for LLM Agents



Project 3A translates the Project 2A architecture into human-AI interactions with Language Model-based AI agents. The central question is **how a conversational (chatbot-style) agent**

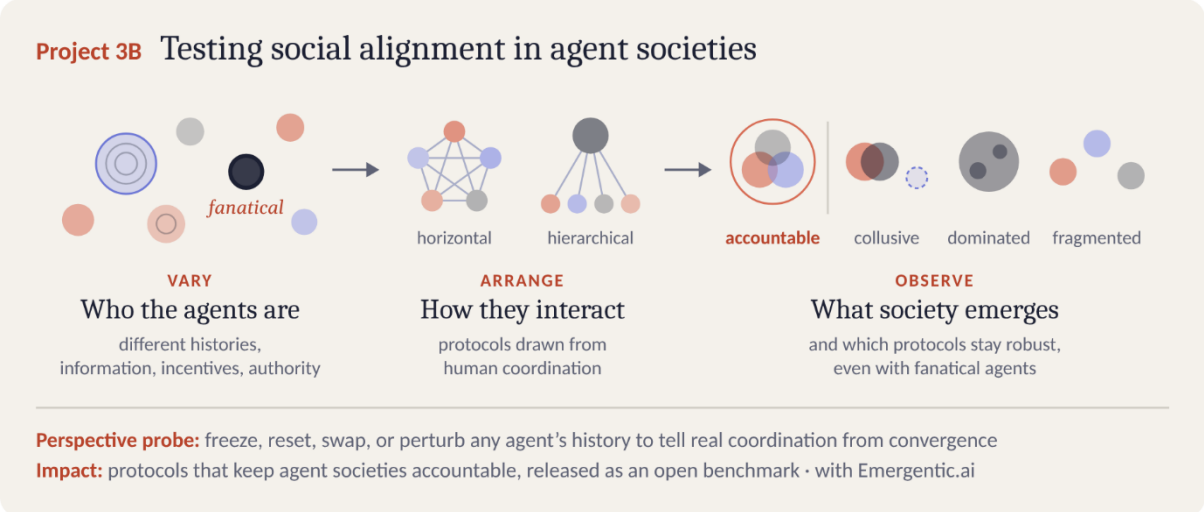
forms and revises an explicit model of the person with whom it interacts over long horizons. Memory retrieval asks what information should be stored and retrieved; the perspective architecture asks how accumulated history should condition the interpretation of new evidence, and how much that evidence should be allowed to revise the accumulated state. In this sense, the problem is not simply one of retaining more information, but of maintaining a history-dependent organization through which new information is interpreted.

The perspective architecture was designed to be model-agnostic, since the slow perspective latent operates separately from the fast policy-optimization layer. It can therefore function as an external wrapper around existing LLMs without requiring modification of the base model itself, and can operate in parallel with retrieval-based memory systems such as RAG. Retrieval supplies episodic or factual context, while the perspective layer maintains a slower, inspectable, history-dependent representation of the user or alignment target that conditions how such evidence is interpreted and incorporated over time. Updates to this state are regulated by confidence, contradiction, repetition, and evidence quality, while response generation remains separated from the pathway that writes persistent state. Provenance and uncertainty remain inspectable, and the state can be frozen, swapped, reset, or otherwise intervened upon for causal evaluation.

Here, the immediate AI alignment-relevant question is the **stability-plasticity tradeoff**: *can a persistent AI resist false overwrite from transient input while still adapting to genuine change?* Initial comparisons will include retrieval-based memory, static profiles, and unconstrained evolving profiles. Relevant failure modes include stale user models, contradictory evidence, state poisoning, sycophantic accommodation, and cases where an agent's own influence on a user becomes confused with independently held preferences.

This project is intended first as a scientific and alignment test of whether a distinct longitudinal perspective state is necessary and useful. Productization is downstream of that question.

Project 3B: Embodied Multi-Agent Perspective & Distributed Coherence



Project 3B extends Project 2B into embodied multi-agent systems. Each unit enters interaction as a body-grounded, history-bearing agent with its own perspective dynamics. The central question is whether **individual coherence can be maintained while interaction produces genuine reciprocal reorganization**, and whether **repeated coupling gives rise to persistent organizational structure at the level of the collective**.

Initial experiments will examine mutual modeling under partial observability, partner-specific adaptation, perturbation and recovery, changing interaction topology with and without unit replacement, and whether identical present configurations behave differently after different collective histories. This provides a tractable route to the CEAR Lab's questions of divergence, resonance, role differentiation, and distributed coherence while preserving the individuality of the participating agents.

A separate causal emergence analysis provides a supporting measurement track. Within a single agent's neural network, causal emergence-based analysis localized the dominant temporal organization to the slow perspective latent, while also showing that scalar integration alone is not a reliable measure of learned organization: learning changed the **composition** of whole-to-whole and whole-to-part influence rather than a simple increase in aggregate magnitude. Project 3B can extend this question one level upward, where the parts are themselves adaptive agents.

Supporting work: Pae (2026), *Perspective Latents as an Architectural Condition for Causal Emergence in Active Inference Agents*, arXiv:2607.20708.

---

6. Roadmap & Outputs (for next 18 months)

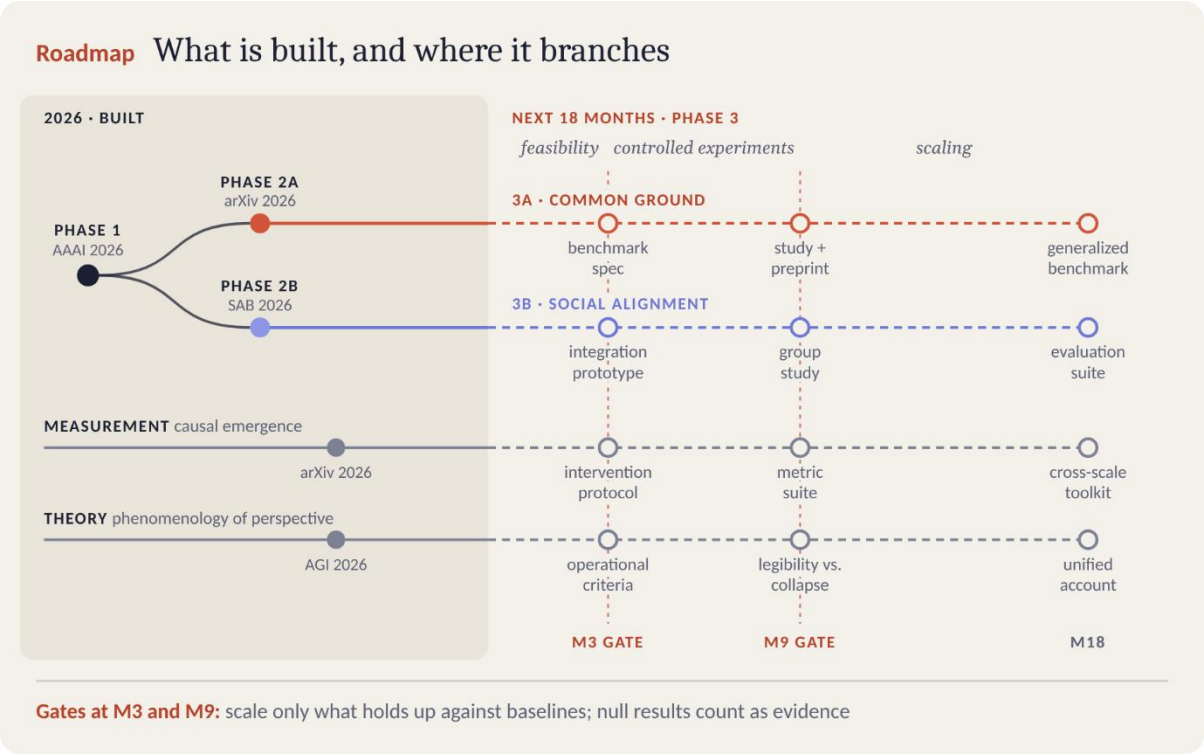
Program Status at the Start of the Roadmap

The core architectural claims of the first two phases of CEAR Lab were empirically tested during 2026. **Phase 1** demonstrated a history-bearing perspective latent that remained temporally persistent and structurally dissociable from fast policy adaptation. **Phase 2** then developed through two complementary projects: **Project 2A** showed that accumulated perspective can reorganize perceptual encoding and regulate its own openness to revision, while **Project 2B** grounded the architecture in interoceptive viability and distinguished perspective formation from conative action-readiness.

Two additional workstreams now run in parallel with the phase-based research. The first is a **measurement and causal analysis track**, which uses causal emergence methods to characterize the temporal whole-part organization associated with the perspective latent and to develop assays that can later be applied at the multi-agent level. The second is a **theoretical track**, which develops the phenomenological and architectural interpretation of perspective as a structural condition for subjectivity and internal coherence. These tracks provide the analytical and conceptual tools through which the empirical projects are interpreted and compared.

The next 18 months focus on **Phase 3** of CEAR Lab program, in which a perspective-bearing

system must interact with another perspective-bearing system. The central objective is to test whether the architectural principles established in Phases 1 and 2 remain informative when perspective becomes relational, i.e. when distinct history-bearing systems must interpret, adapt to, and remain mutually legible to one another.



Phase 3 develops through two parallel projects: **Project 3A**, which extends the architecture to longitudinal human-AI interaction through Language Model-based agents, and **Project 3B**, which extends it to embodied multi-agent systems. Projects 3A and 3B proceed in parallel. The measurement and theory tracks likewise continue in parallel.

Because Projects 3A and 3B enter new technical and empirical domains, the roadmap is milestone-driven. Later-stage scaling decisions will be conditioned on early empirical results and integration feasibility.

Months 1-3: Feasibility, Integration, Assay Design

Can the architecture remain a causally tractable object when extended to relational settings?

The first stage establishes whether the Phase 1-2 architecture can be meaningfully transferred into relational settings and identifies the minimum interventions required to test it causally.

Project 3A

- Implement a minimal model-agnostic perspective layer around existing LLM systems while keeping perspective-state updating separable from response generation and retrieval-based memory. The first benchmark will focus on whether a persistent agent can distinguish enduring changes in a user from transient or anomalous evidence.

- Initial comparisons will include retrieval-only memory, static user profiles, and continuously updated profiles. The perspective state will remain directly inspectable and manipulable through interventions such as freezing, resetting, or swapping state.

Project 3B

- Establish technical and scientific feasibility for integrating CEAR's perspective architecture with an existing embodied LLM multi-agent environment. This stage will identify where persistent agent-specific state can enter the perception-interpretation-action loop, determine which interventions can be performed reproducibly, and define a minimal paired- or small-group interaction paradigm.
- The central question is whether distinct agents can carry different interaction histories into the same present environment in a way that remains experimentally accessible rather than being absorbed into ordinary memory or policy dynamics.

Measurement & Theory

- Causal and dynamical assays will be adapted from the single-agent setting to support direct intervention on relational organization.
- Theory track will refine operational distinctions among perspective, memory, policy, world-model, affect, and conation, and specify what would count as evidence for genuinely history-bearing perspective rather than generic recurrence.

Decision Gate (Month 3)
Progression to full relational experiments will depend on three conditions:

- The perspective state can be manipulated independently of ordinary memory or policy.
- The relevant systems support reproducible replay and perturbation.
- At least one measurable consequence of history survives appropriate baseline controls.

Deliverables

- Project 3A – minimal wrapper and benchmark specification.
- Project 3B – integration prototype and technical feasibility report.
- Shared intervention protocol for freeze/reset/swap/perturb operations.
- Refined operational criteria for history-bearing perspective.

Months 4-9: Controlled Relational Experiments

If so, do history-dependent relational effects actually emerge?

The second stage tests whether history-bearing perspective has causal consequences in sustained interaction.

Project 3A

- Controlled longitudinal dialogue histories to test stable preferences, transient anomalies,

genuine changes, contradictory evidence, misleading statements, return-to-baseline conditions, and state poisoning.

- Central research question asks: “When current context and accessible memory are comparable, does accumulated interaction history produce a distinct, persistent organization that changes how new evidence is interpreted?”
- Primary measures: false overwrite, genuine-change detection latency, stale-state persistence, contradiction resolution, reacquisition after reset, downstream error.
- Causal interventions on the perspective state will test whether observed effects depend on the proposed architecture rather than on accumulated text history alone.

Project 3B

- Controlled paired- and small-group experiments to test whether agents reaching comparable present configurations through different interaction histories subsequently organize themselves differently.
- Initial analysis will examine: asymmetric information and embodied viewpoint; partner-specific adaptation; perturbation and recovery; unilateral conflict or misleading interaction; stable differentiation versus convergence; mutual predictability and interpretive legibility.
- Central research question asks whether different histories remain causally consequential when agents occupy the same shared environment but encounter it from distinct embodied points of view.

Measurement & Theory

- Compare persistence, recovery, reciprocal influence, and whole-part organization across Projects 3A and 3B.
- Theory track will refine the distinction among mutual legibility, resonance, coordination, and collapse into shared state.

Decision Gate (Month 9)
Scaling will proceed only where controlled experiments show robust effects beyond memory and architectural baselines. Null results will be treated as informative evidence about whether the proposed machinery is necessary, and about what additional mechanisms may be required for persistent relational organization.

Deliverables

- Project 3A – controlled longitudinal study and preprint.
- Project 3B – paired/small-group empirical study.
- preliminary metric suite for stability, recovery, mutual predictability, and legibility.

Months 10-18: Scaling, Generalization, and Translation/Application

If so, how far do those effects generalize and scale?

The final stage generalizes the mechanisms supported by the earlier experiments.

Project 3A

- If the perspective layer shows measurable value beyond simpler baselines, the project will expand across longer interaction histories, multiple base language models, retrieval systems, domains, and state-migration conditions.
- A user-facing pilot may be developed to test whether the architecture remains interpretable and stable outside tightly controlled benchmark conditions.
- This stage will also assess potential translational pathways for persistent human-AI systems.

Project 3B

- If relational effects are reproducible in paired or small-group settings, the project will scale toward larger and more heterogeneous agent populations. Candidate extensions include role differentiation, agent replacement, topology changes, damage and recovery, cross-model populations, and strategic social failure modes such as manipulation, coalition formation, polarization, or collective lock-in.
- The central question is whether stable history-bearing organization can emerge at the level of the collective while participating agents remain individually identifiable and reconfigurable.

Measurement & Theory

- Causal emergence and dynamical measures will be extended across scales to test whether whole-level organization persists while components reorganize.
- Theoretical work will integrate perspective formation, perceptual reorganization, embodiment, conation, and relational organization into a unified CEAR Lab account of coherent and mutually legible agency.

Deliverables

- Project 3A – generalized implementation and benchmark.
- Project 3B – embodied multi-agent evaluation suite.
- Empirical manuscripts from Projects 3A and 3B.
- Cross-scale causal analysis toolkit.

---

7. Broader Significance

As frontier AI systems become persistent, agentic, embodied, and increasingly social, the "alignment problem" turns on more than the quality of individual outputs. It raises a deeper question: ***how does internal coherence arise at all?*** Existing approaches can shape external

behavior, but they do not model the internal 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 basis for a more principled understanding of agent alignment as the dynamics through which coherent agency develops across individuals and multi-agent ecologies.

The CEAR Lab program provides a minimal research framework for making these internal dynamics experimentally explicit. Grounded in computational phenomenology, it formulates falsifiable structural hypotheses while addressing what may be the most fundamental form of the alignment problem: **how distinct subjective worlds can come to overlap sufficiently for mutual understanding without collapsing into a single perspective**. 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.

Selected Publications

- Pae, H. (2025). Reflective Analysis on Empirical Theories in Consciousness. Frontiers in Psychology 16, 1571098.
- Pae, H. (2026). Minimal Computational Preconditions for Subjective Perspective in Artificial Agents. Proceedings of the AAAI Symposium Series 8(1), 309-315.
- Pae, H. (2026). Same World, Differently Given: History-Dependent Perceptual Reorganization in Artificial Agents. arXiv:2604.04637.
- Pae, H. (2026). Body-Grounded Perspective Formation and Conative Attunement in Artificial Agents. International Conference on Simulation of Adaptive Behavior, 204-216.
- Pae, H. (2026). Perspective as a Missing Structural Condition for Subjectivity in Active Inference Systems. International Conference on Artificial General Intelligence, 137-143.
- Pae, H. (2026). Perspective Latents as an Architectural Condition for Causal Emergence in Active Inference Agents. arXiv:2607.20708.

Contact Information

Hongju Pae / hjpae@activeinference.institute
Research Fellow, Active Inference Institute

CEAR Lab is an independent research program linking phenomenology, Active Inference, embodied cognition, and computational modeling to develop computational models of phenomenal consciousness and subjectivity as an internalist foundation for agent alignment. Research collaboration and support inquiries are welcome.