Perspectival Control Identity Theory

Jared Moffat

Independent Researcher
jared.e.moffat@gmail.com

Abstract

Perspectival Control Identity Theory (PCIT) proposes a falsifiable, intervention-based identity program for phenomenal consciousness. The central claim is an *a posteriori* identity: phenomenal consciousness is identical to a specific kind of internal control variable, a Perspectival Control State (PCS)—a temporally extended, viability-weighted stream that makes an agent's competing needs comparable and coordinates a coalition of constitutive consumers whose control and learning depend on the stream and whose outputs feed back to shape it. PCIT makes two specific, testable bridge commitments. *Degree* of consciousness tracks how much the PCS stream causally matters for closed-loop viability regulation under intervention. *Content* tracks the decoder-indexed equivalence classes over PCS states and the induced similarity geometry determined by constitutive consumers. Advances in machine learning make this kind of "synthetic phenomenology" experimentally tractable: we can build agents-in-worlds with known internal organization, intervene on proposed PCS implementations and their decoders, and measure downstream effects on behavior and long-horizon viability. The payoff is a research program with concrete invariance and dissociation tests—and a principled scientific foundation for questions about AI moral status, animal sentience, and disorders of consciousness that currently lack one.

Introduction

We are building increasingly sophisticated AI systems with no scientific basis for assessing their phenomenal status. Animal welfare science lacks principled tools for determining which systems are sentient and to what degree. Clinical assessment of consciousness in patients with disorders of consciousness remains primitive, relying on behavioral proxies rather than mechanistic criteria. These are not merely academic problems: they determine which systems warrant moral consideration, how we design and deploy AI, and how we treat non-human animals and patients who cannot report their experience.

What is missing is a *scientifically tractable* theory of phenomenal consciousness—one that yields intervention-ready predictions, engineered falsifiers, and a systematic research program rather than armchair analysis. Perspectival Control Identity Theory (PCIT) is designed to fill that gap. Recent advances in machine learning make the core of this program empirically viable: modern RL systems (Mnih et al. 2015; Silver et al. 2016; Hafner et al. 2020) can serve as synthetic agents-in-worlds with fully known internal organization, allowing direct intervention implementations.

The philosophical starting point is an *a posteriori* identity claim (one discovered empirically rather than derived logically): phenomenal consciousness is identical to a specific organizational kind—a Perspectival Control State (PCS)—in the same way that water is identical to H_2O: an identity nobody predicted a priori, but that becomes unavoidable once the science converges. This identity is not derived from first principles; it is earned empirically by proposing a candidate physical/organizational kind, extracting intervention-ready predictions, and testing them. The explanatory gap between physical descriptions and phenomenal truths (Chalmers 1996; Levine 1983) is treated as an expected epistemic feature of any such identity program, not as a metaphysical barrier: phenomenal concepts can secure reference with first-person authority while failing to reveal the underlying nature of what they refer to (Loar 1990; Papineau 2002; Balog 2009; Stoljar 2005). The remaining metaphysical question—whether this is a standard scientific identity or requires additional psychophysical laws—is not settled by the empirical program PCIT proposes, and can be left open. What matters is establishing the right empirical foundations.

PCIT proposes a substrate-agnostic control architecture in which a PCS is a temporally extended, viability-weighted control stream broadcast to a feedback-coupled set of downstream consumers (systems for action selection, attention/gating, learning, and memory). The theory makes two specific bridge commitments: degree tracks how much the PCS stream causally matters to closed-loop viability regulation, and content tracks decoder-defined equivalence classes over PCS states and the induced similarity geometry. The core empirical bet is that these mappings will be robust under targeted interventions—supporting stabilized biconditionals within a specified architecture class and making identity the best explanation.

When PCIT links PCS to degree and content in a way that (i) explains the structural regularities of experience and (ii) survives interventions designed to break mere correlation, the dialectical landscape changes. At that point, insisting that PCS "merely correlates" with experience requires

a further story: what additional explanatory or causal role does experience play beyond the PCS organization already doing the control-theoretic work?

The paper proceeds as follows. I motivate PCS by isolating the control problem faced by viability-constrained agents with multiple competing needs under resource limits. I then develop the formal architecture: the viability integrator, constitutive consumers, and decoder-indexed content. The two identity bridges are stated explicitly as testable commitments. The remainder identifies the synthetic phenomenology research program, pressure tests, and comparisons to nearby theories.

Perspectival Control States

PCIT begins with a working hypothesis: phenomenal regularities—first-person noticeable and commonly accepted structural features of conscious experience—are the fingerprint of an organizational solution to a specific control problem. Consider an adaptive agent with many welfare-relevant variables (energy, damage, temperature, social threat) and real scarcity (limited bandwidth, compute, memory). Such an agent must repeatedly answer a shifting question: *what matters to me right now?* Water matters more when dehydrated; predator cues matter more under threat; long-run plans matter more when immediate constraints are stable.

A normative sketch makes the pressure concrete. Let $z_t \in \mathbb{R}^K$ be a vector of welfare variables and let $V(z)$ be a viability utility: high within a viable band, steeply falling near constraint violations. Agents are incentivized to learn and model how an action a will produce welfare change $\Delta z(a)$. A local approximation yields

$$V(z_t + \Delta z(a)) \approx V(z_t) + \nabla V(z_t)^\top \Delta z(a). \qquad (1)$$

The gradient $\nabla V(z_t)$ functions as a *stakes* or *urgency* vector: the marginal importance of each welfare dimension *at this moment*. Any system that must trade off heterogeneous needs under scarcity requires some mechanism that turns predicted multi-dimensional consequences into a comparable, live priority signal.

Scarcity then forces coordination across subsystems. The same stakes weighting must guide action selection, attention and compute allocation, and learning/memory updates. If these subsystems optimize against mismatched priorities, behavior fragments: policies pursue one need while attention and learning service another, producing avoidable viability failures.

A natural architectural response is a compact, shared control interface: a low-bandwidth *what-matters-now* stream that keeps downstream subsystems aligned while the world and internal state change. This is a design pressure, not a theorem: for multi-need control under scarcity, architectures that implement a shared, viability-weighted control currency should be stable attractors.

A *Perspectival Control State* (PCS) offers a formal solution to this control problem: a compact internal stream that integrates (i) a "thick present" (what is currently going on), (ii) model-based expectations (what is likely to happen next and what actions afford), and (iii) welfare/viability feedback (what is good/bad/urgent for the agent). The PCS stream functions as a *control currency*: multiple downstream processes condition on it in real time when selecting actions, allocating attention/compute, and gating learning and memory.

PCS is not a dictator. Downstream processes still compute their own policies and updates. PCS shapes the decision space: what is considered, what is prioritized, what gets learned, what gets stored, and which competing control modules win arbitration, given the agent's current stakes under scarcity.

Why a PCS is not merely a "useful latent"

Many networks learn internal states useful for prediction or control. PCS is narrower: it is a shared, viability-weighted coordination interface whose influence is mediated by a constitutive consumer coalition and its feedback knot. Four architectural constraints do the separating work.

Shared access: multiple downstream processes condition on x_t online. *Viability-weighting*: x_t carries urgency/valence structure tied to welfare, not merely sensory summaries. *Coordination under scarcity*: the same stream influences action selection, attention/compute allocation, and learning/memory gating, keeping them aligned when control resources are limited. *No effective bypass*: if a private pathway routes the relevant urgency and coordination signals around x_t, then x_t is not doing the PCS job.

A homeostatic explorer in gridworld

Consider a simple "organism" in a gridworld. It moves and interacts with tiles. Hazards cause damage; movement burns energy; food restores energy; water restores hydration; cold zones drain heat.

The agent receives exteroceptive observations o_t (tiles, objects, local cues) and interoceptive signals tracking welfare variables $z_t \in \mathbb{R}^K$ (energy, hydration, tissue integrity, temperature, toxin load). World dynamics update z_t as a consequence of action and contact ("step onto fire $\rightarrow$ tissue integrity drops"; "run $\rightarrow$ energy drops").

Early in training, the agent is mostly reactive: it stumbles into hazards and fails to anticipate cold zones. Over time it becomes predictive: it learns which sensory patterns forecast future viability loss. "Claws might be nearby" becomes urgent because it predicts tissue damage; "this corridor leads to cold" becomes urgent because it predicts heat loss. Repeated viability hits become teaching events; learning promotes distal cues into early warnings; early warnings prevent emergencies.

On our account, PCS is the device that makes such competence usable by the whole agent. When the agent learns to encode "what matters now" into a shared stream that many subsystems read, action selection, attention allocation, and learning/memory updates stay aligned under resource limits. Interventions on the PCS stream—clamping specific dimensions, injecting noise, severing feedback pathways from particular consumers—should produce systematic, predictable changes in long-horizon viability and coordination behav-

ior, not merely local perturbations. This is what makes the architecture experimentally tractable.

The viability integrator and its inputs

Let $x_t \in \mathcal{X}$ denote the candidate PCS stream. PCIT posits a learned *viability integrator* $\mathcal{I}_\theta$ that constructs x_t directly:

$$x_t := \mathcal{I}_\theta(\underbrace{h_t}_{\text{thick present}}, \underbrace{m_t}_{\text{predictive state}}, \underbrace{z_t}_{\text{welfare context}}, \underbrace{x_{t-1}}_{\text{persistence}}). \tag{2}$$

Here h_t is a short history window; m_t is a learned predictive state (world/self expectations, affordances, counterfactual forecasts); z_t is the welfare vector; and x_{t-1} provides controlled temporal continuity (a flowing present rather than isolated frames).

The integrator is not merely a compressor. Its job is to synthesize a viability-weighted stream for guided control: variations in x_t should track distinctions that matter for the agent's continued viability and that help a coalition of downstream consumers coordinate the allocation of scarce control resources under time pressure.

Viability feedback as the teaching signal and flywheel

The defining claim is that $\mathcal{I}_\theta$ is trained by viability feedback. Let $V(z)$ be a viability function (high within viable ranges, sharply decreasing near constraint violations) and write $v_t := V(z_t)$. A scalar teaching signal uses change in viability,

$$r_t := v_{t+1} - v_t, \tag{3}$$

or any shaped variant preserving the same normative role. Because x_t conditions multiple consumers, changing x_t changes downstream control, which changes future trajectories, which changes v_t. Thus θ is under direct pressure: if $\mathcal{I}_\theta$ fails to route urgency and coordination information into x_t, closed-loop viability collapses.

Backpropagation path. Let ϕ collect parameters of downstream consumers and let the agent optimize expected discounted viability:

$$J(\theta, \phi) := \mathbb{E}\left[\sum_{t \geq 0} \gamma^t r_t\right]. \tag{4}$$

In standard RL implementations (Sutton and Barto 2018; Sutton 1988), updates are driven by a temporal-difference error

$$\delta_t := r_t + \gamma U_\psi(x_{t+1}) - U_\psi(x_t), \tag{5}$$

where U_ψ estimates expected future viability from x_t. Integrator updates follow the chain of influence

$$\Delta\theta \;\propto\; \delta_t \, \nabla_\theta x_t \,(\text{via downstream consumers conditioned on } x_t). \tag{6}$$

This is the flywheel: viability hits create teaching events; learning makes the system sensitive to distal predictors; distal predictors prevent future hits; the PCS stream becomes increasingly about the agent's world and needs in an action-guiding way.

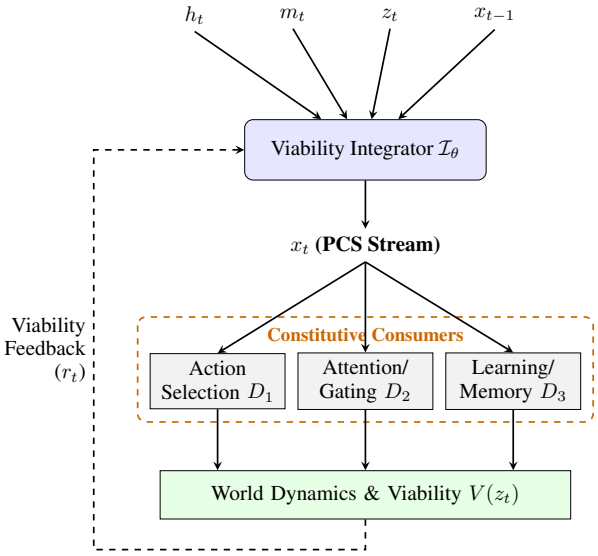

Figure 1: **The PCS architecture:** A viability integrator $\mathcal{I}_\theta$ constructs the shared PCS stream from present, predictive, and welfare states. This stream coordinates downstream constitutive consumers, whose outputs shape environmental dynamics and generate the viability feedback that trains the integrator.

Grounding the viability signal. In engineered agents, the training setup supplies the grounding: the designer specifies z_t as the optimization target. In organisms, grounding is supplied by physiology: homeostatic setpoints, damage signals, and reinforcement machinery provide welfare-relevant error signals. PCIT requires only a stable source of welfare-relevant error capable of training the integrator and its consumers over time.

Consumers, decoders, and decoder-indexed content

PCS is defined by how it is *used*. Let $\{C_i\}_{i=1}^n$ be downstream *consumers* of x_t. Each consumer implements a causal *decoder* that maps PCS into a control-relevant output:

$$y_t^i := D_i(x_t, h_t^i), \qquad h_{t+1}^i := U_i(h_t^i, x_t, o_{t+1}), \tag{7}$$

where h_t^i is the consumer's internal state and y_t^i may be an action proposal, arbitration/gating signal, learning-rate modulation, memory write key, or report-like behavior. These decoder pathways are what make PCS *do work* in the control loop.

A key constraint is coordinate-independence. Two implementations can realize the same control organization while encoding x_t in different internal bases. PCIT therefore individuates content by *within-agent control consequences*, not by raw PCS coordinates.

Constitutive consumers and the subject boundary

PCIT formally defines the "subject" as a control coalition organized around x_t. A downstream subsystem counts as constitutive only if it is (i) welfare-relevant and (ii) feedback-

entangled with PCS. Let $v_t := V(z_t)$ and fix horizon H. Define the discounted viability functional

$$W_{t:t+H} := \mathbb{E}\left[\sum_{k=0}^{H-1} \gamma^k \, v_{t+k}\right], \qquad 0 < \gamma \leq 1. \quad (8)$$

A consumer C_i is *constitutive* only if it satisfies both conditions below.

(V) Viability-relevant dependence. There exists a PCS intervention (Pearl 2009) that changes the consumer's decoded output and thereby changes expected future viability:

$$\exists\, \tilde{x} \;\text{ s.t. }\; \mathbb{E}[W_{t:t+H} \mid \mathrm{do}(x_t = \tilde{x})] \neq \mathbb{E}[W_{t:t+H}]\,. \quad (9)$$

(F) Feedback entanglement. The consumer's outputs must causally shape future PCS beyond a transient blip:

$$\exists\, \kappa \leq H,\ \exists\, \tilde{y} \;\text{ s.t. }\; P\big(x_{t+\kappa} \mid \mathrm{do}(y_t^i = \tilde{y})\big) \neq P(x_{t+\kappa})\,. \quad (10)$$

Passive "spectator" readouts typically fail (F) and are excluded. Let $\mathcal{K}$ be the set of consumers satisfying (V) and (F). The "subject" adds no further ontology: it is the maximal viability-relevant, strongly coupled coalition organized around a shared control currency.

Decoder-indexed equivalence classes (content-types). Fix $\mathcal{K}$ and define the decoder profile map $\mathcal{D}(x; \mathbf{h}_t) := \big(D_i(x, h_t^i)\big)_{i \in \mathcal{K}}$. This induces an equivalence relation on PCS space:

$$x \sim_{\mathbf{h}_t} x' \quad \text{iff} \quad \mathcal{D}(x; \mathbf{h}_t) = \mathcal{D}(x'; \mathbf{h}_t). \quad (11)$$

Phenomenally: two PCS states are the same content-type at time t if and only if they make no difference to any constitutive consumer's control output. Content is fixed by the consuming pathways inside the welfare-relevant feedback knot, not by an observer's readout. Any invertible reparameterization of x that preserves constitutive decoder profiles preserves content.

Decoder-induced similarity geometry. Equip each decoder output space with a control-relevant metric d_i and define the induced distance

$$d(x, x' \mid \mathbf{h}_t) := \sum_{i \in \mathcal{K}} w_i \, d_i\big(D_i(x, h_t^i),\, D_i(x', h_t^i)\big)\,. \quad (12)$$

Phenomenally: this distance encodes the similarity structure of experience as seen from the inside—which experiences are close or far apart in the agent's own discrimination-and-control space. In humans, phenomenal similarity structure is empirically accessible: subjects can reliably report relative similarity ("A is closer to B than to C"), and those judgments predict discrimination thresholds, confusion matrices, and generalization gradients (Shepard 1987; Kriegeskorte, Mur, and Bandettini 2008). PCIT's content bridge predicts that these observable similarity relations match the decoder-indexed distances induced by constitutive consumers. This yields a concrete test schema: estimate human similarity structure from behavior, estimate PCS geometry from candidate neural/functional decoders, and evaluate fit and intervention stability.

Degree as viability-weighted causal mattering. PCIT distinguishes two notions. *Engagement* is whether x_t is currently online as a shared interface, in the sense that changes propagate through constitutive decoders to reorganize arbitration, attention, learning gates, and action selection. *Regulatory efficacy* is whether that influence improves expected viability over a horizon. Engagement and efficacy can dissociate: a PCS stream can be vividly engaged while steering maladaptively, and some viability can be sustained by habitual controllers even when PCS engagement is reduced. What matters for PCIT's degree bridge is the viability-weighted causal influence of PCS over a horizon, with engagement as the primary pathway but not a guarantee of beneficial control.

The Two Identity Bridges

PCIT makes two specific, testable commitments linking PCS structure to phenomenal structure. These are the falsifiable core of the theory.

Bridge for content. Phenomenal character/content corresponds to decoder-indexed PCS state-types: the equivalence classes $[x_t]_{\sim_{\mathbf{h}_t}}$ induced by the constitutive consumers (11), together with the decoder-induced similarity geometry (12). Content is not fixed by raw coordinates of x_t, but by which distinctions in x_t make a difference to the control outputs of the constitutive coalition.

Bridge for degree. Degree of phenomenal presence corresponds to how much the PCS stream causally matters to the agent's closed-loop regulation over a horizon—viability-weighted causal influence on coordinated downstream control. Operationally, this is measured by comparing $\mathbb{E}[W_{t:t+H}]$ under targeted interventions on x_t against baseline, weighted by the downstream importance of the resulting control changes. Degree is measured by a family of such intervention-based mattering metrics; PCIT is committed to the existence and testability of such measures, not a unique formalization.

Why access asymmetry is expected. PCS is an internal control interface built to drive the constitutive coalition. First-person acquaintance is constituted by being among the consumers whose coordinated dynamics the stream is designed to regulate: the system is "reading" and "acting from" its own shared interface. Third-person observers can model, decode, and intervene on the PCS state but occupy an external vantage point: they are not among the constitutive consumers inside the welfare-relevant feedback knot. The familiar first-/third-person asymmetry is not an extra metaphysical problem added on top of the identity claim; it is the expected difference between a state playing its role inside the agent's closed-loop control economy and a state described from outside.

What PCIT does not claim. PCIT does not promise a third-person entailment of "what it is like." Whether the identity is a standard scientific reduction or requires additional psychophysical laws is not settled by the empirical program proposed here, and can be left open. What

PCIT claims is that PCS is the right *level of description* at which phenomenal consciousness becomes scientifically tractable—sufficient to ground a rigorous, intervention-based research program and to yield actionable criteria for the practical questions that motivate the theory.

Research Program

PCIT is meant to be testable, not merely interpretable. In synthetic systems we can *build* candidate PCS architectures, *read and intervene* on internal streams and decoder pathways, and *match* agents that differ only in whether the PCS constraints are satisfied. The result is a program with clean internal knobs and engineered falsifiers.

Evidence ladder. *Humans are the calibration set*: report is one access channel, but the core constraints are the richer psychophysical structures that accompany report—similarity judgments, discrimination thresholds, confusion matrices, metamers, masking and attention dissociations, and graded loss/recovery across sleep and anesthesia (Brown, Lydic, and Schiff 2010). These fix a target similarity structure for content and a target profile of control-relevance for degree. *Synthetic agents serve as a forcing function*: they let us build candidate PCS architectures, intervene directly on x_t and consumer decoders, and test the invariances and dissociations that the bridges predict.

The synthetic goal is not an incorrigible "machine qualia" verdict. A robot can report that it "sees red," but verbal confession is not the right test. Instead, PCIT uses task families that expose internal similarity structure directly: forced-choice discriminations, graded generalization, confusability under noise/occlusion, metamers induced by controlled input transforms, and systematic dissociations under bandwidth throttling or consumer decoupling. Treat "report" as just another consumer—it counts only if it is inside the viability-relevant feedback knot and improves long-horizon regulation when accurate. The synthetic question is: does a PCS-like architecture exhibit the predicted invariances and dissociations, and does the decoder-indexed geometry function as the unique intervention-stable handle that explains its similarity structure and control syndromes? A negative answer is an engineered falsifier; a positive answer is evidence that the organizational kind and bridge principles are coherent, buildable, and nontrivially constraining.

Pressure tests and possible falsifiers

Unused-bits. If large subspaces of x_t can be randomized without changing any constitutive decoder outputs or coordinated downstream control, PCIT predicts no content change. If content reports track those unused coordinates, PCIT fails.

Codec/player swap. If one can reparameterize or recode x_t while preserving the full constitutive decoder profile and the coordinated control syndrome, PCIT predicts content invariance. If content changes with mere coordinate changes, PCIT fails.

No-consumers. If a candidate PCS stream is present but has no constitutive consumers, PCIT predicts no experience. If a "phenomenal" syndrome persists without constitutive consumers, PCIT fails.

Split-consumers. If the constitutive coalition splits into two weakly coupled groups with separate feedback knots, PCIT predicts split or degraded unity. If unity remains intact despite a clean coalition split, PCIT fails.

Reparameterization invariance. If two different implementations realize the same decoder-indexed partition and induced geometry (up to isomorphism) but yield systematically different content-geometry signatures, PCIT fails. The converse also holds: if implementations with different raw coordinate systems but identical decoder profiles produce identical content signatures, that is positive evidence for the theory.

External probes vs. in-loop modules. A passive readout should not become constitutive. A module wired into the feedback knot in a viability-relevant way should become constitutive. If passive probes alter content, or in-loop modules do not, PCIT fails.

Ontogeny I: stakes-first maturation. In systems trained under viability pressure, PCIT predicts a characteristic developmental ordering: early learning primarily tunes *stakes*, while later learning stabilizes a compact shared stream that multiple consumers can use for coordinated control. The decoder-indexed partition should become sharper over training, and the induced similarity structure should become more stable. If task competence increases while the PCS stream remains geometrically unstable, PCIT is undercut.

Ontogeny II: welfare-relevant incorporation. When the environment makes a new predictive feature reliably welfare-relevant, PCIT predicts it becomes "present" in the PCS sense only when it modulates stakes and reorganizes coalition control through x_t. If agents exploit new welfare-relevant structure while the PCS stream and constitutive decoder profile remain unchanged, that indicates an effective bypass and counts against the PCS claim.

Synthetic phenomenology: three research tracks

A serious synthetic program has three tracks. *Discovery* in existing agents avoids stipulation and reveals what modern training actually produces. *Construction* by design creates clean knobs and controlled internal variables. *Matched comparisons* isolate what PCS contributes beyond generic capacity or broadcasting. Ontogeny—how PCS and its consumer coalition co-develop over training—cuts across all three.

Track 1: PCS hunting in existing RL architectures. Start with successful RL systems (actor–critic variants, recurrent policies, world-model agents, transformer-based agents) and ask: is there a compact internal stream that (a) multiple downstream subsystems condition on in real time, (b) is viability-weighted, (c) coordinates action, attention/compute, and learning/memory gating, and (d) lacks an effective bypass? The aim is to identify candidate shared interfaces and test whether their causal role matches the

bridge commitments: degree should covary with viability-weighted causal mattering over a horizon, and content-like similarity structure should align with decoder-indexed partition/geometry rather than raw coordinates.

Track 2: PCS by design. Engineer architectures that intentionally implement the PCS constraints: impose an explicit shared bottleneck that is the only route by which welfare-weighted priorities can jointly influence multiple consumers; train on multi-need tasks with explicit welfare variables and nonlinear penalties outside viable ranges; then verify the shared stream is doing the coordinating work rather than functioning as a convenient latent while private pathways carry urgency. This track yields isolated experimental knobs—stream bandwidth, consumer coupling strength, welfare tradeoff structure—that are hard to guarantee in purely discovered architectures.

Track 3: Matched PCS vs. non-PCS comparisons. Construct matched agent pairs equalized for parameter count, training data, and task performance, but differing in whether they satisfy the PCS architectural constraints: a shared bottleneck versus private specialized pathways; a single shared stream versus multiple separate streams; strong versus weak consumer coupling. Then compare bridge-shaped signatures: does viability-weighted causal mattering of the shared stream track graded presence-like syndromes under load or throttling? Does decoder-indexed geometry predict discriminability and invariances under coordinate changes? The engineered falsifiers provide clean pass/fail targets.

From synthetic success to brain-facing hypotheses. If these tracks converge in synthetic agents, we gain principled hypotheses for biological brains: a shared, viability-weighted control stream with multiple real-time consumers and no effective bypass, such that degree tracks viability-weighted causal mattering and content tracks decoder-indexed partition/geometry. One interesting hypothesis is that part of the viability training signal is carried by neuromodulatory systems. In mammals, dopamine is a plausible candidate for a fast, signed teaching signal tracking whether outcomes are better or worse than expected (Schultz, Dayan, and Montague 1997), while other neuromodulators (norepinephrine, acetylcholine, serotonin) may tune arousal, attention, and plasticity in ways that reweight which concerns become urgent (Aston-Jones and Cohen 2005; Yu and Dayan 2005). Neuromodulators are not the PCS itself; they help train and regulate the viability integrator machinery that produces PCS.

Comparing PCIT to Other Approaches

PCIT borrows what is empirically useful in nearby frameworks, then adds the bridge commitments needed to target *phenomenal* consciousness with intervention-based tests.

Global-workspace proposals (Baars 1988; Dehaene and Naccache 2001) capture the shared-interface insight well: when information becomes broadly available, control becomes more flexible, coordinated, and reportable. PCIT accepts this insight but treats "global availability" as at best a marker for *access*. The relevant shared stream in PCIT is a scarcity-constrained control currency trained under viability pressure—its function is to make competing concerns comparable and actionable. That functional role, not broadcast per se, is what PCIT ties to phenomenal presence and character.

Predictive processing and active inference (Friston 2010; Clark 2013) treat self-regulation via predictive models and error signals as central. PCIT treats this as plausible background engineering and expects PCS to be implementable in such architectures. But predictive systems generate many local errors that never matter to the agent's cross-need control economy. Phenomenal presence and character, on PCIT, track the specific viability-weighted interface that actually arbitrates priorities and gates downstream control. Prediction can help construct and update PCS; it is not the identity candidate.

Higher-order and report-based accounts (Block 1995; Dehaene 2014) identify a robust pattern: conscious episodes tend to guide deliberate control, memory, and report. PCIT accepts the pattern but relocates the explanans. The key variable is not "having a higher-order state" or "being reportable" in abstraction; it is being routed into and used by the constitutive consumer coalition. Report is one consumer among others, correlating with consciousness when it sits inside the same viability-relevant feedback knot and dissociating when it does not.

Integrated information approaches (Tononi 2004; Oizumi, Albantakis, and Tononi 2014) start from phenomenological constraints and seek an intrinsic structural measure. PCIT shares the aim of respecting phenomenal structure but differs critically in bridge strategy. Rather than elevating an intrinsic quantity as the primary explanans, PCIT proposes a candidate organizational kind (PCS) and demands *intervention-stable* bridges. Crucially, PCIT admits engineered falsifiers that IIT does not: one can build a system with high integrated information that fails every PCS test, and vice versa—a PCS-organized system with low Φ. This makes the theories empirically distinguishable, not merely philosophically different. Degree is tied to causal mattering in closed-loop regulation; content is tied to decoder-indexed geometry; both are invariant to arbitrary internal coordinate choices and representational bases.

Conclusion

PCIT is an identity-style bet: phenomenal consciousness is a specific organizational kind, not an epiphenomenal glow on top of structure. The candidate kind is a shared, viability-weighted control interface (PCS) that coordinates an agent's downstream control economy under scarcity. The theory makes two separable, testable bridge commitments. Degree tracks how much the PCS stream causally matters to closed-loop regulation over a horizon. Content tracks the decoder-indexed partition/geometry induced by constitutive consumers inside the loop. Both claims are tested by intervention and invariance.

The first-person/third-person asymmetry stops looking metaphysically spooky on this account: PCS is built to be used by the agent's own control coalition, and acquaintance is constituted by playing the inside-the-loop role. "Subject"

talk is a label for the maximal viability-relevant feedback-entangled coalition; boundary questions become substantive only when that coalition genuinely splits or expands. The remaining metaphysical question—standard identity or psychophysical laws—is not settled by the empirical program proposed here, and can be left to the armchair. What matters is that PCIT provides the right empirical foundations.

PCIT succeeds if PCS becomes the unique intervention-stable handle that unifies the major regularities: graded presence via causal mattering, structured content via decoder-indexed geometry, and unity/perspective via a single control coalition. It fails if these links break under pressure tests. Synthetic agents are the forcing function: in silico we can build the candidate kind, intervene directly, and test invariances that are hard or impossible in vivo.

The practical stakes are significant. A successful PCIT program would provide principled, empirically grounded criteria for assessing AI moral status—replacing intuition and behavioral mimicry with measurable architectural facts. It would give animal welfare science a mechanistic basis for determining which systems are sentient and to what degree. And it would offer clinical medicine better tools for assessing consciousness in patients who cannot report their experience. These are not distant applications; they are the reason getting the science right matters urgently.

References

Aston-Jones, G.; and Cohen, J. D. 2005. An Integrative Theory of Locus Coeruleus–Norepinephrine Function: Adaptive Gain and Optimal Performance. *Annual Review of Neuroscience*, 28: 403–450.

Baars, B. 1988. *A Cognitive Theory of Consciousness*. Cambridge University Press.

Balog, K. 2009. Phenomenal Concepts. In McLaughlin, B.; Beckermann, A.; and Walter, S., eds., *The Oxford Handbook of Philosophy of Mind*. Oxford University Press.

Block, N. 1995. On a Confusion about a Function of Consciousness. *Behavioral and Brain Sciences*, 18(2): 227–247.

Brown, E. N.; Lydic, R.; and Schiff, N. D. 2010. General Anesthesia, Sleep, and Coma. *New England Journal of Medicine*, 363: 2638–2650.

Chalmers, D. J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Clark, A. 2013. Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. *Behavioral and Brain Sciences*, 36(3): 181–204.

Dehaene, S. 2014. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Viking.

Dehaene, S.; and Naccache, L. 2001. Toward a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework. *Cognition*, 79(1–2): 1–37.

Friston, K. 2010. The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience*, 11: 127–138.

Hafner, D.; Lillicrap, T.; Ba, J.; and Norouzi, M. 2020. Dream to Control: Learning Behaviors by Latent Imagination. In *International Conference on Learning Representations (ICLR)*.

Kriegeskorte, N.; Mur, M.; and Bandettini, P. 2008. Representational Similarity Analysis—Connecting the Branches of Systems Neuroscience. *Frontiers in Systems Neuroscience*, 2: 4.

Levine, J. 1983. Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly*, 64(4): 354–361.

Loar, B. 1990. Phenomenal States. *Philosophical Perspectives*, 4: 81–108.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; et al. 2015. Human-Level Control through Deep Reinforcement Learning. *Nature*, 518(7540): 529–533.

Oizumi, M.; Albantakis, L.; and Tononi, G. 2014. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology*, 10(5): e1003588.

Papineau, D. 2002. *Thinking about Consciousness*. Oxford University Press.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition.

Schultz, W.; Dayan, P.; and Montague, P. R. 1997. A Neural Substrate of Prediction and Reward. *Science*, 275(5306): 1593–1599.

Shepard, R. N. 1987. Toward a Universal Law of Generalization for Psychological Science. *Science*, 237(4820): 1317–1323.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; et al. 2016. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature*, 529: 484–489.

Stoljar, D. 2005. Physicalism and Phenomenal Concepts. *Mind & Language*, 20(5): 469–494.

Sutton, R. S. 1988. Learning to Predict by the Methods of Temporal Differences. *Machine Learning*, 3: 9–44.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition.

Tononi, G. 2004. An Information Integration Theory of Consciousness. *BMC Neuroscience*, 5: 42.

Yu, A. J.; and Dayan, P. 2005. Uncertainty, Neuromodulation, and Attention. *Neuron*, 46(4): 681–692.