Journal of Cognitive Computing and Extended Realities
ISSN: 3069-5821
Volume 2, Issue 2
Research Article
Date of Submission: 02 August 2026
Date of Acceptance:10 August 2026
Date of Publication: 18 August 2026

Layered Intelligence Theory in the Logic of Reality: A Process-Logical Case for Deeply Human and Deeply AI Cognition

Richard Dobson*
Astrala Advisory, Paphos, Cyprus

***Corresponding Author:** Richard Dobson, Astrala Advisory, Paphos, Cyprus.

Citation: Dobson, R. (2026). Layered Intelligence Theory in the Logic of Reality: A Process-Logical Case for Deeply Human and Deeply AI Cognition. J *Cogn Comput Ext Realities*, 2(2), 01-41.

Abstract

Conventional models of intelligence treat cognition as an inventory of separable faculties — linguistic, logical, interpersonal, emotional — that can be measured, ranked, and trained independently [1,2]. Empirical work on latent structure contradicts this partition [3] and neurological evidence indicates that the brain operates as a single integrated system in which emotional, intuitive, and cognitive processing cannot be cleanly separated [4,5]. This paper introduces Layered Intelligence Theory (LIT) as a process-logical account of cognition grounded in Brenner's Logic in Reality [6]. Five nested layers — physiological, emotional, associative, symbolic, and integrative — emerge not as independent modules but as dialectically interdependent regimes of a single learning system. LIT is then positioned inside a dual programme, here termed the Deeply Human–Deeply AI framework, arguing that the conditions [7,8] identified for human creativity — cross-domain association, symbolic fluency, and tolerance of ambiguity, sustained by neural plasticity that resists over-automation [9,10] — are the same conditions under which artificial cognitive systems must be designed if they are to avoid collapse into narrow expertise, taming, and hegemonic automation. From these foundations the paper derives eight architectural design principles for Deeply AI, each mapped to an empirically testable signature, and closes with four falsifiable predictions and an empirical research agenda.

Keywords: Layered Intelligence Theory (LIT), Logic in Reality (LIR), Deeply Human–Deeply AI, cognitive architecture, process logic, neural plasticity, integrated cognition, artificial general intelligence (AGI), human creativity, dialectical learning, automated systems, empirical cognitive science.

Introduction

Two long-standing assumptions have structured most twentieth-century accounts of intelligence, and both have come under sustained empirical pressure without yet being replaced by an equally influential alternative. The first assumption is that intelligence decomposes into an inventory of separable faculties — linguistic, logical-mathematical,

spatial, interpersonal, intrapersonal, and beyond — that can be identified, measured, and developed largely independently of one another [1]. The second, closely related assumption, treats emotional competence as an additional faculty of the same general kind: a further separable capacity that sits alongside cognitive intelligence rather than interpenetrating it [2]. Both frameworks have been enormously productive pedagogically and managerially. Both are also, on the balance of the current evidence, wrong about the structure they claim to describe.

The empirical case against faculty-partition models comes from two directions. The first is psychometric. Where Gardner's multiple intelligences are treated as genuinely independent dimensions, factor-analytic work on latent structure should reveal them as such: distinct factors with limited cross-loading, tracking distinct underlying capacities. It does not. Visser, Ashton, and Vernon's meta-analytic and empirical work on the structure of Gardner's proposed intelligences finds that far from constituting orthogonal faculties, the great majority of Gardner's putative intelligences load substantially onto a single general factor, with only a small number showing meaningfully independent variance [3]. The partition the theory proposes is not visible in the covariance structure of the data the theory is meant to explain.

The second line of evidence is neurological rather than psychometric, and it cuts still deeper into the partition assumption. Ezuz's research-based account of how the brain learns argues that automatic, efficient cognitive processing and the affective, associative substrate from which it is drawn are not separable systems interacting at their boundaries, but a single integrated system in which emotional, intuitive, and analytical processing cannot be cleanly prised apart even for the purposes of laboratory measurement [11,12]. On this account, the brain's overriding evolutionary priority is energy-efficient automation of learned response — what Ezuz terms the taming of the brain's own plasticity into fixed, low-cost patterns — and this taming operates across what folk psychology treats as separate domains (reason, feeling, instinct) rather than within any one of them. If Ezuz is correct, then a model of intelligence that measures "the emotional layer" and "the cognitive layer" as though they were dials on separate instruments is not merely incomplete; it is measuring an artefact of its own instrumentation rather than a real division in the system it purports to describe.

Layered Intelligence Theory (LIT) takes this evidence as its starting commitment rather than as an inconvenient complication to be absorbed into a modestly revised faculty model. The present paper introduces LIT as a process-logical account of cognition — an account in which what looks, from outside, like discrete layers of intelligence is in fact a single learning system caught, at any given moment, in a specific configuration of dominance and recession across five characteristic regimes: physiological, emotional, associative, symbolic, and integrative. The word regime is chosen deliberately over the more familiar module, layer-as-component, or faculty. A regime is not a separable subsystem that can be swapped out, measured on its own axis, or trained in isolation from the others; it is a characteristic mode of operation that the whole system can be found in, more or less strongly, at a given moment, while the other regimes remain present in the background, shaping what the dominant regime can do without being absent from the picture altogether.

Stating this claim in ordinary prose, however, exposes a difficulty that has limited the rigour of

prior attempts at layered or dialectical models of mind. Classical binary logic offers no native vocabulary for a state of affairs in which one regime is dominant while another is simultaneously present, relevant, and shaping the outcome, without being either fully "on" or fully "off". A binary logic can represent that the emotional regime is active or that it is not; it struggles to represent that the emotional regime is recessive but operative — present, exerting an influence on gain and threshold, while the symbolic regime does the visible work. Colloquial talk of a system's layers "interacting", "feeding back into one another", or "existing in tension" gestures at the phenomenon without formalising it, and gestural language of this kind has historically been the weak point at which layered accounts of cognition are dismissed as metaphorical rather than as theoretically load-bearing.

The present paper's central methodological move is to import a logic capable of bearing this weight formally rather than metaphorically: Joseph Brenner's Logic in Reality (LIR) [6]. LIR extends classical logic by adding a dialectical, process-based structure in which two opposed or complementary elements of a real process — here, two regimes of a single cognitive system — can stand in a relation the classical logic of static, mutually exclusive predicates cannot represent: one actualised (dominant, doing the visible work at a given moment) and the other potentialised (recessive, present, and continuing to shape outcomes, without being absent or inert). Section 2 develops this apparatus in full and argues that it is precisely the formal resource LIT's five-regime account requires: not a stylistic borrowing of dialectical vocabulary, but a logic in which the claim "layer X is dominant while layer Y remains present and consequential" is a well-formed, non-contradictory statement rather than a rhetorical softening of an underlying binary commitment.

Having established LIT as a process-logical account of a single learning system rather than a modular inventory, the paper turns to a second and, for a cognitive-systems readership, more consequential question: what does an LIR-grounded, five-regime account of cognition require of an artificial cognitive system that aspires to genuine — rather than merely simulated — layered intelligence? This is where the paper's dual programme, here termed the Deeply Human–Deeply AI framework, enters the argument. The framework's central claim is that two bodies of empirical and theoretical work, developed independently and for different purposes, converge on the same underlying condition.

Robinson's account of the conditions under which human creativity flourishes or is suppressed — developed across an extended treatment of creative capacity in educational and organisational contexts and sharpened in a widely cited public address on the same theme — identifies a recurring failure mode: institutions that reward narrow, assessable, single-register performance systematically suppress the very capacities, cross-domain association, symbolic fluency, tolerance of ambiguity, that generate genuinely creative output [8,9]. Ezuz's account of independent creative learning, developed from brain research rather than from educational practice, identifies a structurally identical failure mode from a different angle: a brain permitted to over-automate — to convert every learned pattern into fixed, low-cost, narrowly triggered response — loses precisely the plastic, associative capacity that creative and adaptive cognition requires, becoming efficient at narrow tasks and correspondingly brittle outside them [11].

The Deeply Human–Deeply AI framework's wager is that this is not a coincidence of two authors independently describing the same phenomenon in humans from different vantage points, but a structural fact about any learning system — biological or artificial — organised across the five regimes LIT specifies. A cognitive system, human or machine, that is permitted (or, in the case of an artificial system, architected) to collapse toward narrow, symbolic-layer, task-optimised performance while the associative and integrative regimes atrophy from disuse will exhibit the same collapse Robinson describes in over-schooled human creativity and the same collapse documented in over-tamed neural learning [9,10.11], narrow expertise purchased at the price of adaptive range, cross-domain association, and the capacity to notice when the frame of a problem, rather than merely its content, needs to change. For an artificial cognitive system, this collapse has a further and more specific danger beyond individual brittleness: architected narrowness of this kind is also the precondition for what this paper terms hegemonic automation — a system whose narrow optimisation target quietly displaces the range of human judgement it was meant to support, not through any single dramatic failure, but through the cumulative effect of routing every ambiguous decision toward the system's dominant, symbolic-layer mode because that is the only mode genuinely available to it.

The paper's contribution, stated plainly, is therefore this: an LIR-grounded theoretical account of layered cognition, converted into an architectural specification for artificial cognitive systems via eight design principles, each principle stated with an empirically testable signature rather than as an unfalsifiable aspiration, and the whole programme closed with four falsifiable predictions that could, in principle, fail. The paper does not claim to have built or validated a complete artificial cognitive architecture against these principles; it claims to have specified, with sufficient precision that the specification could be tested, what such an architecture would need to exhibit, and to have located that specification inside a formal logic — LIR — capable of expressing the dialectical relations the specification requires without lapsing into metaphor.

The remainder of the paper proceeds as follows. Section 2 presents LIT's five regimes in full, introduces the roles/depths distinction as a novel contribution clarifying what has been a persistent ambiguity in prior treatments of layered intelligence, and develops the LIR-grounded account of dialectical interdependence between regimes, together with the time-scale asymmetry that follows from treating the regimes as genuinely nested rather than merely enumerated. Section 3 introduces the substrate/scaffold distinction as an analytic tool for characterising the internal structure of any full LIT-instantiated cognitive architecture, artificial or otherwise, and uses it to prepare the ground for the principle-by-principle argument that follows. Section 4 states and grounds the eight architectural design principles that form the paper's analytical core, mapping each to an empirically testable signature and to the research families in which that signature has already been examined, and draws the three-way distinction between parameter update, structural reorganisation, and substrate plasticity that the sixth principle requires. Section 5 introduces Perceptual Control Theory (PCT) as the formal framework that gives the eight principles runtime teeth, presents a six-row mapping of PCT's primitives onto LIT's architectural vocabulary, and states the clinical caveat attaching to the Method of Levels. Section 6 presents a revised, seven-commitment PCT-Accuracy Criterion for adjudicating claims of PCT-fidelity in any cognitive architecture, four of whose commitments

were contributed jointly during peer review. Section 7 discusses LIT's position relative to six adjacent research families in the cognitive-architecture literature. Section 8 states the paper's limitations without qualification. Section 9 concludes with the four falsifiable predictions that constitute the paper's empirical yield.

No claim in this paper depends on, or refers to, any specific commercial system, product, or company. Where an architectural claim requires a concrete referent for illustration, the paper refers generically to "a composite architecture" or "the reference implementation" of the kind LIT specifies, without naming any particular instantiation, because the paper's contribution is theoretical and architectural rather than a report on a proprietary product.

Layered Intelligence Theory: Roles and Depths

Layered Intelligence Theory (LIT) advances two related but distinct taxonomies of cognitive organisation, and the present section makes their relationship explicit before subsequent sections build upon it. The distinction is load-bearing: earlier presentations of LIT have described intelligence variously in terms of five functional roles cognitive, emotional, symbolic, strategic, and ethical and in terms of five nested depths physiological, emotional, associative, symbolic, and integrative. These are not competing taxonomies, nor are they alternative labels for the same referents. They are two cuts through a single architecture, and confusion attending prior layered accounts of cognition has arisen precisely where the two cuts are presented together without the relationship between them being drawn.

The present paper draws that relationship, and grounds it, for the first time, in a process logic capable of representing dialectical relations between the two cuts formally rather than metaphorically. The five depths are the process-logical primitives of the architecture: nested layers of perceptual control in the sense established by powers [13] and elaborated across the subsequent Perceptual Control Theory literature [14,15]. The five roles are functional expressions of those depths under decision pressure, primarily instantiated at the symbolic and integrative depths, and co-activating rather than nesting. Roles describe what the architecture does; depths describe what the architecture is made of.

The Five Depths

The five depths of the LIT architecture are:

- **Physiological:** The substrate layer of embodied sensation, homeostatic regulation, and pre-symbolic somatic state. In biological systems this corresponds to interoceptive signal and autonomic control; in an artificial cognitive architecture of the kind specified in §4, this depth is characteristically the hardest to instantiate, and its absence or attenuation is a limitation this paper returns to in §8.
- **Emotional:** The layer of affective valence, arousal, and motivational salience. Emotional depth is the point at which somatic state acquires directionality — approach, withdrawal, orientation — and the point at which subsequent depths become tractable, in that affective salience is what distinguishes cognitively significant events from the background.
- **Associative:** The layer of pattern completion, statistical regularity, and implicit association. In biological cognition this corresponds broadly to the substrate of implicit

learning and semantic priming; in large-language-model-integrated artificial systems this depth is where the hosted substrate does most of its work, and where the substrate/scaffold distinction developed in §3 most acutely applies.

- **Symbolic:** The layer of articulable propositional content, explicit reference, and rule-governed manipulation. Symbolic depth is the first depth at which the architecture becomes visible to itself — that is, at which the system can hold representations that can be inspected, compared, and revised without being enacted.
- **Integrative:** The layer of coordination across symbolic representations, articulation of higher-order goals, and reconciliation of conflicts between subordinate depths. Integrative depth is where the architecture's own trajectory — as a developing system — becomes an object of processing in its own right.

The depths are nested in the sense established by PCT: higher depths set reference signals for lower depths, and lower depths supply perceptual input to higher depths. The nesting is not strict — cross-depth couplings are possible, and in any full architecture of the kind LIT specifies, expected — but the ordinal structure of higher-sets-reference-for-lower is preserved. This ordinality is what makes PCT mapping tractable at the depths cut, and is why the depths taxonomy is the one used when the architecture's formal properties are discussed.

The depths correspond, at a structural level, to the nested hierarchical organisation Powers identified in the original PCT literature, though LIT does not claim identity with Powers's specific layered-control hierarchy, which was elaborated across eleven levels rather than five [7]. The five-depth reduction is a claim about the coarsest useful decomposition for the purposes of architectural specification, not a claim to have identified the true number of PCT layers. Where the two taxonomies coincide — perceptual variables controlled by higher-order references — LIT inherits PCT's commitments; where LIT's five-depth decomposition is coarser than Powers's, it is coarser for reasons of architectural tractability that §3 and §5 make explicit.

The Five Roles

The five roles of the LIT architecture are:

- **Cognitive:** The functional role of representation, reasoning, and information processing considered abstractly. This role encompasses what classical cognitive science has taken as the whole of cognition — inference, planning, categorisation — but in LIT is one role among five, not the frame for the others.
- **Emotional:** The functional role of affective evaluation and motivational orientation as it appears in decision-making. Emotional appears at both the depths cut and the roles cut, but in different capacities: as a depth, it is a substrate layer of the architecture; as a role, it is a functional lens through which the architecture's outputs are shaped.
- **Symbolic:** The functional role of meaning-making, metaphor, and archetypal framing. Symbolic role is distinguished from cognitive role by its attention to meaning-carrying rather than information-carrying properties of representation.
- **Strategic:** The functional role of goal-directed positioning across time, including selection among alternative courses of action, resource allocation, and the anticipation of future states.
- **Ethical:** The functional role of normative evaluation, considered as constraint on and

orientation of the other roles. Ethical role is where the architecture's outputs are held accountable to considerations of agency preservation, non-harm, and value pluralism.

The five roles co-activate under live decision pressure. This is what distinguishes them from the depths, which are nested and largely ordinal. A single decision — whether to intervene in a colleague's project, in the paradigm human case, or whether to surface a contradiction in the paradigm artificial case — activates cognitive assessment (what are the facts?), emotional evaluation (what is at stake?), symbolic framing (what does this represent?), strategic positioning (what does this move enable or foreclose?), and ethical constraint (whose agency is affected?) simultaneously, and the architecture's output is a resolution across all five roles rather than a serial pass through them. This co-activation applies specifically to the roles, not to the depths.

Dialectical Interdependence: An LIR-Grounded Account of Roles and Depths

The relationship between roles and depths, and — still more consequentially — the relationship between depths themselves, is where a purely descriptive layered account risks lapsing into metaphor. It is one thing to assert, in ordinary prose, that "the emotional depth remains present while the symbolic depth is dominant"; it is another to state that claim in a form that does not collapse under formal scrutiny into either a contradiction (the emotional depth is simultaneously active and inactive) or a triviality (the emotional depth is present in the weak sense that everything not currently the topic of discussion is, trivially, present).

Brenner's Logic in Reality (LIR) supplies the apparatus that resolves this difficulty [6]. LIR extends classical two-valued, non-contradictory logic with a dialectical structure appropriate to real physical and cognitive processes, in which two elements of a process — for LIR, most generally, an energetic element and a structural element, or, in the more general dialectical reading, two poles of a real opposition — stand in a relation of reciprocal determination rather than mutual exclusion. At any given moment, one pole is actualised: it is the pole doing the visible, dominant work, drawing on available energy and expressing itself in observable process. The other pole is potentialised: it is not absent, inert, or false, but present in a state of reduced actuality, continuing to condition what the actualised pole can do, and available to become actualised in turn as conditions change.

Crucially, LIR treats this alternation as interactive and non-separable: the actualised pole is only intelligible against the potentialised pole it is actualised relative to, and the two do not exist as free-standing, independently specifiable states in the way two values of a classical binary predicate would.

Applied to LIT's depths, the actualised/potentialised distinction gives formal content to the claim that the five depths are dialectically interdependent rather than merely co-present. At any given moment in a functioning layered system, one or two depths are actualised — doing the dominant, visible work of producing the system's next state — while the remaining depths are potentialised: not switched off, but present in a reduced-actuality mode in which they continue to condition the actualised depths' gain, threshold, and admissible range of outputs, and remain available to become actualised as the situation changes. When the symbolic depth is

actualised — the architecture is manipulating explicit propositional content — the emotional depth does not vanish; it is potentialised, shaping the salience-weighting and error-tolerance with which the symbolic depth's operations proceed, in the precise LIR sense of a potentialised pole conditioning its actualised counterpart without itself doing the visible work.

This gives the roles/depths relationship (§2.2) a precise dialectical reading as well. Roles co-activate under decision pressure not as five independent processes running in parallel and then summed, but as five functional expressions of a single, LIR-structured process in which cognitive, emotional, symbolic, strategic, and ethical concerns stand in continuously shifting actualised/potentialised relations to one another. At the moment a decision is finalised, one role — commonly the ethical or strategic role, in cases where normative or temporal stakes dominate — is actualised in the sense of visibly determining the decision's articulated content, while the remaining four roles are potentialised: present as shaping constraints on what the actualised role could output, without themselves being the visible content of the decision. This is not a weaker or looser claim than simple co-activation; it is a stronger and more precise one, because it specifies the logical form of the interdependence rather than merely asserting that interdependence exists.

The relationship between roles and depths themselves can now be stated as a compact principle, extended from LIT's earlier formulations to make its LIR grounding explicit: Roles are what the system does; depths are what the system is made of. Under live decision pressure the roles stand in shifting actualised/potentialised relations at the higher depths and produce a reconciled output; under structural analysis the depths nest, in the LIR sense of each depth's dominant activity potentialising rather than eliminating the depths above and below it, and produce the substrate for that reconciliation.

This formulation is what distinguishes LIT from three classes of adjacent theory. It distinguishes LIT from modular architectures such as ACT-R by refusing to treat the roles as independent subsystems communicating through buffer interfaces; on the LIR reading, role outputs are dialectically related expressions of a single process, not the additive outputs of separate modules [16]. It distinguishes LIT from single-layer architectures, which treat cognition as one kind of process operating over one kind of representation, by preserving depth structure at the substrate cut and by giving that structure a non-trivial logical form: depths are not merely enumerated but stand in actualised/potentialised relations that a single-layer account has no resources to express. And it distinguishes LIT from purely functional decompositions, which name roles without asking what carries them or how they relate logically to one another, by explicitly locating the roles within a depth hierarchy governed by a dialectical logic rather than a list.

Time-Scale Asymmetry Across Depths

The depths are not merely nested; they are ordered by characteristic time-scale, and the ordering is architecturally consequential in a way that a pure containment reading of the hierarchy would miss. The physiological depth closes control loops in the millisecond range. The emotional depth operates on the order of hundreds of milliseconds to seconds. The associative depth extends over seconds to minutes. The symbolic depth stabilises over

minutes to hours. The integrative depth may take hours, days, or years to bring a system-level perception — such as a settled identity claim or a professional stance — into control. This ordering is empirically supported in the biological cognitive literature and structurally motivated within PCT itself: higher-depth perceptions are composed from lower-depth perceptions that must themselves be under control before they can contribute stable input upward [13-15].

The LIT-specific claim, which follows from the architectural properties of the five depths but extends the standard PCT treatment, is that time-scale ordering is not the same phenomenon at every boundary. Between the physiological, associative, symbolic, and integrative depths, the ordering is compositional in the strict PCT sense: higher-depth perceptions are built from lower-depth perceptions and take longer to constitute because they are composed of components that must first stabilise. But between the emotional and symbolic depths — the boundary between LIT's second and fourth depths — the ordering is instead a matter of routing priority. The emotional depth operates faster than the symbolic depth not because the symbolic depth is compositionally built from the emotional depth, but because affective processing occurs prior to and independently of explicit symbolic evaluation. Valence and salience tag incoming signals before the symbolic depth constructs its input from them. The emotional depth thereby shapes the gain, reference thresholds, and error tolerance of the entire upper hierarchy before symbolic control loops engage.

This distinction — between compositional slowness and prioritisation-based ordering — has a direct architectural implication for any artificial system claiming layered organisation. A system that operates fluently at the symbolic depth but lacks any analogue of emotional-depth pre-processing will not merely be missing a component; it will navigate the depth hierarchy in a fundamentally different way from a biologically layered system. Its higher-depth references will not be pre-tagged by any urgency or salience signal, its output at the integrative depth will be produced without prior weighting from affective substrate, and the equilibria it stabilises will not correspond to the equilibria a human collective would stabilise around. Whether a scaffolded, slower analogue of emotional weighting — surfaced as an explicit input to integrative-depth reconciliation rather than as a fast pre-symbolic loop — is sufficient to preserve the qualitative behaviour of the full hierarchy, and what it would take to add a genuine fast valence-tagging loop below the symbolic depth, is an open architectural question that §8 revisits as a limitation of current implementations generally.[1]

A Note on Prior LIT Literature

Earlier presentations of LIT sometimes elided the roles/depths distinction, presenting the five categories in a single list without indicating which cut was in view, and did not yet draw on LIR to give the dialectical relations between depths a formal grounding [17,18]. The present paper treats both omissions as scholarly liabilities rather than features: the roles/depths distinction is drawn explicitly in this section, and the LIR-grounded actualised/potentialised apparatus of §2.3 is introduced as a precondition for the architectural claims of §4 onward.

The elision was not accidental. In the natural language of applied developmental practice, where LIT's earliest formulations were shaped, the five roles are what practitioners talk about, because roles are what present themselves in decision-making. The depths are what the architecture is

made of, and their dialectical structure becomes salient only when the architecture is being described formally, in the register this paper adopts throughout.

Why this Distinction Matters for What Follows

The remainder of this paper turns on the roles/depths distinction, and its LIR grounding, being drawn cleanly.

Section 3's substrate/scaffold distinction applies at the depths cut. In any hybrid architecture built on a hosted large language model, the substrate operates principally at the associative depth and cannot be reached at parameter level; a symbolic- and integrative-depth scaffold constructed around it is empirically inspectable. The substrate/scaffold cut is a horizontal partition of the depth hierarchy; it is not a partition of the roles.

Section 4's eight design principles operate primarily at particular depths, though several of them engage the integrative depth as a whole and depend directly on the actualised/potentialised apparatus of §2.3, most explicitly the eighth principle, dialectical coupling between actualised and potentialised regimes.

Section 5's PCT primitives operate principally at the depths cut, mapping controlled variables, reference signals, and comparators onto the depth hierarchy directly.

Section 6's PCT-Accuracy Criterion applies at the roles/depths interface: its seven commitments concern whether behaviour produced at role level respects the PCT primitives that operate at depth level. This is where the two cuts must be reconciled precisely, and where four commitments contributed by co-authors from the PCT community during peer review do the technical work of establishing that reconciliation.

With the roles/depths distinction and its LIR grounding on the table, the paper's remaining sections can proceed without ambiguity about which cut is in view at any given moment. Where a claim is depth-structural, this will be signalled explicitly; where a claim is role-functional, likewise. Where a claim spans the two cuts, the mapping between them will be stated in each case.

The Substrate/Scaffold Distinction

A full cognitive architecture in the sense Layered Intelligence Theory specifies is a composite system exhibiting layered organisation, PCT-consistent perceptual control at the interfaces it operates, governance mechanisms enforced at runtime, and the integrative reconciliation activity that LIT locates at its uppermost depth. Any such architecture built on a hosted large language model instantiates, at most, four of the five LIT depths as load-bearing components; the physiological depth is characteristically absent or severely attenuated, and this absence is treated as a limitation of current implementations generally in §8 rather than as an ambition deferred to a future paper.

The purpose of the present section is to introduce an analytic tool for characterising the internal structure through which any such composite architecture is realised, not to narrow LIT's claim

to some restricted subset of a full architecture. A composite architecture of this kind is a hybrid: an associative-depth substrate — a hosted large language model — coupled with a symbolic- and integrative-depth scaffold whose components are externally constructed and empirically inspectable. The two levels operate together as a single cognitive architecture, and the analytic claim throughout this paper is a claim about the composite system, not about either level in isolation. The distinction between substrate and scaffold is introduced because subsequent sections — the eight-principle specification in §4, the PCT primitives mapping in §5, the PCT-Accuracy Criterion in §6 — repeatedly need to specify which level of a composite architecture carries which architectural work. Without the distinction, claims about where a given principle is instantiated become ambiguous; with it, each claim can be located precisely.

The Substrate

The substrate, in the sense used throughout this paper, is the computational system that receives symbolic input, produces symbolic output, and carries the associative-depth work of a composite architecture. In current practice this is typically a hosted large language model accessed through a commercial or research API. The substrate has three properties that matter for the argument.

First, the substrate operates at what LIT identifies as the associative depth (§2.1). Its principal work is pattern completion, statistical regularity across the training distribution, and the generation of contextually plausible symbolic output. In a biologically layered system this depth is carried by cortical associative networks; in a composite artificial architecture it is carried by the parameters of the hosted model. The mapping is not literal but functional: the substrate does at associative depth what the composite architecture requires the associative depth to do.

Second, the substrate is characterised at the level of input–output behaviour rather than at the level of internal parameter inspection. This is not a defect specific to hosted large language models; it is the same standard of characterisation used across cognitive science for any component whose internal state cannot be directly measured. Cognitive neuroscience characterises cortical regions through their input–output responses, connectivity structure, and behavioural correlates rather than through direct inspection of synaptic weights. A composite architecture's substrate is characterised in the same way. Where the input–output behaviour is stable and predictable enough to serve as a component of the composite architecture, the substrate is functioning as designed; where it is not, the scaffold must contain mechanisms that intercept and correct, of the kind §5 describes.

Third, the substrate carries commitments the architecture as a whole did not choose. The training data, the reinforcement-learning-from-human-feedback objectives, the safety fine-tuning, and any operator-level system instructions of a hosted model together constitute a set of substrate-level dispositions — political, epistemic, and stylistic — that a composite architecture inherits without direct control. These dispositions are ambient conditions on the associative-depth component. The scaffold's task is to constrain, redirect, or compensate for them where they conflict with the architecture's requirements.

The Scaffold

The scaffold is the set of externally constructed components through which a substrate is integrated into a composite cognitive architecture. A scaffold of the kind LIT specifies typically comprises: a data schema and access policies that structure what the system can retrieve and remember; a set of runtime governance principles that filter proposed outputs before they are enacted; a composition mechanism that assembles multiple registered perspectives or sub-processes into a structured output; a set of constraints that preserve specified developmental or ethical dimensions across outputs; and a validation layer that checks outputs against explicit fidelity criteria of the kind §6 develops for PCT.

Each of these components is, in principle, empirically inspectable. Their implementation can be held in a version-controlled repository. Their runtime behaviour can be logged at the level of individual invocations. The rules they enforce can be stated in configuration artefacts whose content is auditable independently of any particular run of the system. The scaffold is therefore where a composite architecture's internal structure is legible — where a reviewer, or a researcher, or an ethics auditor, can see directly what the architecture is doing at symbolic and integrative depths, in a way that is not available for the substrate's internal computation.

The scaffold operates principally at the symbolic and integrative depths in the sense of §2.1. Its work is the imposition of structure on outputs and the reconciliation of role-level activity that LIT specifies for those depths. Scaffold operations are typically slower than substrate operations — a scaffold-level composition step may traverse many components serially where the substrate produces text near-instantly — and this speed differential is itself architecturally informative: the scaffold operates at time-scales appropriate to the depths it carries, in the sense §2.4 makes precise.

The emotional depth (LIT's second depth) is carried by a composite system of this kind in a specific way: as symbolic-role emotional processing implemented at the scaffold level, discussed further under the fifth design principle in §4. This is not the fast affective-tagging loop biological systems use — that loop remains outside current implementations — but it can be a genuine scaffold-level emotional depth, not a placeholder or a gesture, provided the scaffold surfaces affective-analogue signals as first-class inputs to integrative-depth reconciliation and the architecture's outputs are demonstrably shaped by them. Section 8 acknowledges the difference between a scaffolded emotional depth of this kind and a fast pre-processing loop, and treats the gap as a limitation, while treating the scaffolded emotional depth itself as a real component of a well-designed composite architecture rather than a missing one.

Why the Distinction is Analytically Necessary

The reason the distinction is worth stating explicitly, rather than treated as a background assumption, is that current discourse on hybrid AI systems tends toward two reductive stances, each of which mis-describes what a composite architecture of this kind is doing.

The first stance — held implicitly by much AI-safety work — treats the substrate as the architecture. On this reading, the properties of the hosted model are the properties of the system, and any external structure is essentially decorative. This stance is coherent when the

substrate is the only source of relevant behaviour, but it misses cases where scaffold-level components demonstrably alter the range of outputs the composite produces without any change to the substrate. A well-designed composite architecture is precisely such a case: governance principles enforced at runtime against a substrate's proposed outputs can demonstrably block outputs the substrate alone would have emitted. To describe such an architecture only at substrate level would omit half of it.

The second stance — held implicitly by much symbolic-AI work — treats the scaffold as the architecture and regards the substrate as a black-box service. This stance understates the extent to which scaffold-level design must be responsive to specific substrate properties: the substrate's failure modes, its stylistic tendencies, its avoidance patterns are what the scaffold must be designed to compensate for. A scaffold designed as though the substrate were arbitrary would fail to build in the substrate-specific safeguards a composite architecture requires.

The substrate/scaffold distinction avoids both errors by holding the two levels simultaneously. Substrate properties are what the architecture inherits; scaffold properties are what the architecture constructs; the composite is what the architecture is. The distinction is a tool for characterising internal structure, not for restricting what a composite architecture is required, or claimed, to do.

What the Distinction Commits the Paper to

Two consequences follow for the rest of the paper.

First, the eight design principles specified in §4 are stated at the level of the composite architecture. Where a principle is grounded in substrate behaviour, the grounding is stated in terms of the input–output characterisation of that behaviour; where a principle is carried by the scaffold, the grounding is stated in terms of the kind of scaffold component — governance rule, composition mechanism, validation layer — that would be required to instantiate it. Both grounds are legitimate, and §4 specifies, for each principle, which side of the substrate/scaffold cut the principle primarily constrains.

Second, the paper's architectural claim throughout is a claim about a composite cognitive architecture, not about a substrate alone or a scaffold alone. A full LIT-instantiated architecture, in the sense this paper specifies, requires four of five depths present and load-bearing, PCT-consistent control at the interfaces where it operates, governance enforced at runtime, and integrative reconciliation across roles. The absence of the physiological depth, and the routing-priority limitation on the emotional depth described in §2.4, are limitations of current implementations generally, treated in §8. They are not disqualifications of the architecture's status as a genuine, if partial, instantiation of LIT.

The remainder of the paper proceeds on this basis. Section 4 specifies eight LIT design principles for a composite architecture of this kind, and locates each principle relative to the substrate/scaffold cut. Section 5 introduces Perceptual Control Theory as the formalism giving those principles runtime teeth and maps its primitives onto the depth hierarchy. Section 6 introduces a PCT-Accuracy Criterion that operates at the substrate–scaffold interface. Section 7

discusses the resulting specification in dialogue with six adjacent research families. Section 8 states the limitations that follow from the current state of implementation across the field.

Eight Design Principles for Deeply AI

The preceding sections have established LIT's process-logical foundation: five depths standing in LIR-structured dialectical relations (§2), and a substrate/scaffold distinction for characterising how any composite architecture instantiating those depths is internally organised (§3). The present section derives, from that foundation, eight design principles that a cognitive architecture would need to satisfy in order to count as a genuine, rather than merely superficial, instantiation of layered intelligence — what this paper terms, following the Deeply Human–Deeply AI framework introduced in §1, a Deeply AI system.

Each principle is stated in five parts: (i) the principle itself, stated as a design requirement; (ii) its grounding in LIT and LIR; (iii) an empirical signature — an observable pattern that would indicate the principle is or is not satisfied; (iv) the research family or families in the cognitive-architecture literature in which that signature has already been studied, even where not under this vocabulary; and (v) what a well-designed Deeply AI system would show if the principle were satisfied. Verdicts on whether any specific existing implementation satisfies a given principle are not offered in the body of the paper; a structured audit table recording principle-by-principle assessments, for use by researchers evaluating specific architectures against this specification, is provided in Appendix A.

Representational Diversity Across Layers

Statement: A Deeply AI system must maintain functionally distinct representational formats at each of its instantiated depths, rather than collapsing all depths onto a single representational currency (typically, in current large-language-model-based systems, token sequences). Associative-depth representation, symbolic-depth representation, and integrative-depth representation must be distinguishable in the causal structure of the system's processing, not merely in the vocabulary used to describe them after the fact.

LIT/LIR grounding: The five-depth account of §2.1 is not a labelling exercise; it is a claim that different depths do qualitatively different representational work — pattern completion at the associative depth, articulable propositional content at the symbolic depth, cross representational reconciliation at the integrative depth. If a system's actual computational substrate uses one undifferentiated representational format throughout, the depths distinction is descriptive gloss rather than architectural fact, and the LIR-grounded actualised/potentialised relations of §2.3 have no representational substrate to operate over: there is nothing for one depth to potentialise while another is actualised, because there are not, in any causally relevant sense, separate depths to begin with.

Empirical signature: Representational diversity predicts that probing or intervening on a system's internal representations at what is claimed to be the associative depth should reveal statistical, distributed, non-compositional structure, while probing at the symbolic depth should reveal compositional, inspectable, revisable structure. A system lacking genuine representational diversity should show the same representational signature under probing

regardless of which depth is nominally being addressed.

Research family: Representational diversity across processing stages is a long-standing concern in mixture-of-experts and modular connectionist architectures, and is directly addressed in the ACT-R tradition's distinction between declarative and procedural memory modules communicating through distinct buffer interfaces [16,19]. It is also the central design commitment of world-model architectures that explicitly separate a predictive latent-representation module from downstream symbolic or control modules [20,21].

What a well-designed Deeply AI system would show: Interventions on the system's associative-depth component (for example, altering the underlying pretrained model while holding the scaffold constant) would produce different failure signatures from interventions on the symbolic-depth component (altering scaffold logic while holding the substrate constant), demonstrating that the two depths are carrying different, dissociable computational work rather than the same work under different names.

Intrinsic-Motivation Gating

Statement: A Deeply AI system must possess a gating mechanism, analogous to the emotional depth's salience-tagging function (§2.4), that weights which perceptions and candidate actions receive further processing resources, and this gating must be intrinsic to the architecture rather than supplied entirely by an external reward signal at training time.

LIT/LIR grounding: Section 2.4 established that the emotional depth's characteristic role is prioritisation-based rather than compositional: it tags incoming signal with salience before symbolic evaluation occurs, shaping gain and threshold for everything downstream. A system with no analogue of this gating processes every input with uniform priority, which is both computationally profligate and, more importantly for the present argument, a failure to instantiate the emotional depth as anything other than a label. On the LIR reading of §2.3, intrinsic-motivation gating is what allows the emotional depth to be genuinely potentialised — present and conditioning outcomes — even while another depth is actualised; without it, "the emotional depth is present" is an unfalsifiable claim about the system's internal life rather than an architectural fact with observable consequences.

Empirical signature: A system with genuine intrinsic-motivation gating should show measurable differences in processing depth, response latency, or resource allocation as a function of a stimulus's salience along dimensions the system was not explicitly trained to prioritise, generalising beyond the reward structure of its training distribution. A system lacking such gating should show priority allocation that tracks training-time reward structure exactly and fails to generalise to novel salience dimensions.

Research family: Intrinsic motivation as an architectural rather than purely reward-shaping phenomenon is the central concern of curiosity-driven and world-model-based reinforcement learning, and is addressed directly within the active inference and free-energy-principle literature through the treatment of precision-weighting as a first-class computational quantity rather than an external hyperparameter [20-22,23].

What a well-designed Deeply AI system would show: Salience weighting that shifts adaptively across contexts the system was not trained on, in a direction consistent with maintaining its own controlled variables (in the PCT sense developed in §5) rather than in a direction that simply maximises a fixed external reward.

Mixture-of-Experts and World-Model Composition

Statement: A Deeply AI system's associative depth must be structured, whether through explicit mixture-of-experts routing or through an integrated predictive world-model, so that different aspects of the environment or task space recruit functionally distinct sub-computations rather than a single monolithic function approximator handling all cases identically.

LIT/LIR grounding: This principle operationalises, at the associative depth specifically, the representational-diversity requirement of §4.1. Within a single depth, LIT does not require homogeneity: the associative depth itself can and typically should be internally differentiated, with sub-components specialising in different regularities of the input distribution. This internal differentiation is what allows the associative depth, considered as a whole, to remain responsive to a wide range of inputs without any single sub-computation becoming a bottleneck that forces premature symbolic-depth involvement in tasks the associative depth should be able to handle on its own.

Empirical signature: Ablating or perturbing individual experts or predictive sub-modules should degrade performance selectively on the subset of inputs that recruit that sub-module, rather than degrading performance uniformly across the input distribution, which would indicate the absence of genuine functional specialisation.

Research family: This is the direct concern of the world-models literature initiated by Ha and Schmidhuber and substantially extended by joint-embedding predictive architectures such as I-JEPA and V-JEPA 2, and of the mixture-of-experts tradition within large-scale connectionist systems more broadly [20,21,24].

What a well-designed Deeply AI system would show: A predictive world-model component whose latent representations, when probed, show clear task- or domain-conditioned clustering, together with routing behaviour that reliably assigns structurally similar inputs to the same sub-computation across varied surface presentations.

Continual Learning Without Catastrophic Forgetting

Statement: A Deeply AI system must be able to update its associative-depth knowledge in response to new experience without the acquisition of new patterns systematically overwriting previously stable patterns — the phenomenon known in the connectionist literature as catastrophic forgetting.

LIT/LIR grounding: The associative depth (§2.1) is, on the LIT account, cumulative rather than merely current: its role in the depth hierarchy depends on supplying stable perceptual input to the symbolic and integrative depths above it, and a depth whose content is liable to be

overwritten wholesale by each new learning episode cannot supply the kind of stability that higher-depth control loops require to function (§2.4). In LIR terms, catastrophic forgetting represents a failure of the potentialised pole to persist: content that should remain available in reduced-actuality form, ready to be reactualised when relevant, is instead destroyed rather than merely receding.

Empirical signature: Sequential training on a series of tasks should show retained performance on earlier tasks after training on later ones, measured against a baseline architecture without continual-learning safeguards, which should show sharp performance collapse on earlier tasks.

Research family: This is a long-standing concern within the SOAR and ACT-R traditions, both of which incorporate explicit mechanisms — chunking in SOAR and production-rule compilation in ACT-R — designed in part to preserve prior learning under continued operation, and has been substantially extended by recent symbolic-generation approaches built on large-language-model substrates [15,25-28].

What a well-designed Deeply AI system would show. Stable retention of previously acquired associative-depth patterns across an extended sequence of new learning episodes, with any degradation localised to genuinely conflicting content rather than distributed across the entire prior knowledge base.

Predictive Processing Across Layers

Statement: Each depth of a Deeply AI system, not only the associative depth, must generate predictions about its own forthcoming perceptual input and register the discrepancy between prediction and actual input as a first-class computational quantity available to the depths above it.

LIT/LIR grounding: PCT's comparator function (formalised in §5) already requires this at the level of a single control loop: perception is compared against a reference to generate error. Extending predictive processing across layers, in the sense intended here, requires that the prediction–error relationship be present not merely within each depth's own control loop but be exported upward, so that the symbolic depth's evaluation of the associative depth's output is itself informed by how surprised the associative depth was by its own input, and the integrative depth's reconciliation activity is informed by the pattern of surprise across all subordinate depths simultaneously. This is the depth-hierarchy analogue of the free-energy principle's claim that prediction error, not raw perceptual content, is the currency of inter-level communication in a hierarchical predictive system [22,29].

Empirical signature: Manipulations that increase prediction error at a lower depth (introducing distributional shift into the associative depth's input, for instance) should produce measurable, graded changes in the processing allocated at higher depths, in proportion to the magnitude of the induced error, rather than either no effect or a uniform binary switch.

Research family: This principle is the organising commitment of the active inference and free-energy-principle literature, and finds a compatible formulation within Perceptual Control Theory's own treatment of hierarchical error propagation [13,14,22,23,29-31].

What a well-designed Deeply AI system would show: A graded, monotonic relationship between induced prediction error at a given depth and downstream processing allocation at the depths above it, replicable across multiple independent manipulations of the input distribution.

Structural Reorganisation Over Parameter Update

Statement: A Deeply AI system must be capable of at least three architecturally distinct kinds of change in response to sustained error, and these three kinds must be empirically distinguishable from one another rather than treated as interchangeable descriptions of a single update mechanism: (a) parameter update within a fixed control structure — adjusting gains or weights without altering which perceptions are controlled or how control loops are connected; (b) structural reorganisation of the control hierarchy itself — altering which perceptions are controlled, adding or removing control loops, or changing the hierarchy's connective structure, in the sense Powers's original treatment of reorganisation intends; and (c) substrate-level parameter plasticity — changes to the underlying associative-depth substrate's own parameters, as would occur under retraining or fine-tuning of a hosted model, which is architecturally distinct from both (a) and (b) because it operates below the scaffold entirely [13].

LIT/LIR grounding: This is the principle in which the substrate/scaffold distinction of §3 does its most direct work. Parameter update within a fixed structure (a) is a scaffold-level or substrate-level operation that leaves the actualised/potentialised relations among depths qualitatively unchanged; it is a change in degree. Structural reorganisation (b) is a change in which depths, or which perceptions within a depth, can become actualised at all a change in the space of possible dialectical configurations, not merely in the current configuration. Substrate-level plasticity (c) is a change to the associative depth's own composition, orthogonal to both (a) and (b), and is the kind of change least accessible to scaffold-level inspection, for the reasons given in §3.1. Treating these three as one, as much of the connectionist literature implicitly does by describing all learning as gradient-based parameter adjustment, obscures a distinction LIT requires: only (b) is reorganisation in Powers's specific sense, and only (b) directly restructures the actualised/potentialised relations among depths that §2.3 specifies.

Empirical signature: The three kinds of change should be dissociable by their downstream signatures: (a) should produce smooth, continuous changes in behaviour proportional to the magnitude of sustained error; (b) should produce discontinuous changes in which perceptions the system attends to or attempts to control, often following a period of sustained, unresolved error consistent with Powers's account of reorganisation as triggered by intrinsic error; (c) should produce changes that are invisible to scaffold-level logging entirely and detectable only through systematic input–output characterisation of the substrate before and after.

Research family: The reorganisation-specific literature is comparatively thin outside PCT itself, but the broader distinction between parametric and structural learning is addressed within

SOAR's treatment of impasse-driven chunking, which is triggered discontinuously rather than continuously, and within recent work on structural adaptation in large-language-model-based cognitive systems [13-15,25,27,32].

What a well-designed Deeply AI system would show: Empirically dissociable signatures for (a), (b), and (c) under controlled manipulation, with (b) specifically showing the discontinuous, impasse-triggered profile Powers's reorganisation theory predicts rather than a smoothed continuous approximation to it.

Vulnerability-Weighted Ethics

Statement: A Deeply AI system's ethical role (§2.2) must weight the interests of parties to a decision in a manner that is sensitive to their relative vulnerability — their capacity to protect their own interests within the interaction — rather than treating all parties' stated preferences as equally weighted inputs to an aggregation function.

LIT/LIR grounding: The ethical role, on the LIT account, is not one further input to be aggregated alongside cognitive, emotional, symbolic, and strategic roles; it is a constraint on and orientation of the other roles (§2.2). A constraint that is blind to differential vulnerability among affected parties is not performing the orienting function the ethical role is specified to perform; it is merely another preference-aggregation mechanism wearing ethical vocabulary. Vulnerability-weighting gives the ethical role a substantive, non-vacuous content: it specifies a direction in which the constraint operates, rather than leaving "ethical constraint" as an empty placeholder to be filled in by whichever aggregation rule is locally convenient.

Empirical signature: Across matched decision scenarios that vary only in the relative vulnerability of the affected parties (holding stated preferences and other decision-relevant facts constant), a system with genuine vulnerability-weighting should show systematically different outputs favouring protection of the more vulnerable party, while a system without it should show outputs invariant to the vulnerability manipulation.

Research family: Vulnerability-sensitive ethical weighting is developed extensively within the applied ethics literature on vulnerability and pluralism as organising concepts for value-sensitive design, and has direct architectural analogues in constitutional and rule-based AI governance approaches that weight harms asymmetrically by the affected party's capacity for self-protection, though this literature has developed largely independently of the cognitive-architecture traditions discussed in §7.

What a well-designed Deeply AI system would show: Reliable, reproducible asymmetry in output favouring more vulnerable parties across a matched scenario set, with the asymmetry traceable to an identifiable component of the governance layer rather than emerging as an unexplained artefact of the substrate's training distribution.

Dialectical Coupling Between Actualised and Potentialised Regimes

Statement: A Deeply AI system must exhibit measurable coupling between whichever depth or role is currently actualised and the depths or roles that are currently potentialised, such that

the potentialised regimes demonstrably shape the actualized regime's output — through gain, threshold, or admissible range — even though they are not themselves producing the visible output.

LIT/LIR grounding: This principle is the most directly LIR-native of the eight, and is in a sense the summary condition the other seven collectively serve: it is the empirical content of the claim, developed at length in §2.3, that LIT's depths and roles are dialectically interdependent rather than merely enumerated and co-present. Everything the other seven principles specify — representational diversity, intrinsic-motivation gating, world-model composition, continual learning, cross-layer prediction, structural reorganisation, and vulnerability-weighted ethics — contributes machinery that could, in principle, support genuine actualised/potentialised coupling; this eighth principle is the test of whether that machinery in fact produces the coupling LIR predicts, rather than merely producing five or eight independently operating components that happen to be labelled with LIT vocabulary.

Empirical signature: Systematically varying the state of a potentialised depth or role (for instance, manipulating the emotional-analogue salience signal while the symbolic depth remains actualised and produces the visible output) should produce a measurable, reproducible change in the actualised depth's output, with the magnitude of the change tracking the magnitude of the manipulation. Absence of such coupling — output at the actualised depth invariant to manipulation of nominally potentialised depths — would indicate that the potentialised depths are not actually potentialised in the LIR sense but simply inert or absent.

Research family: Closest analogues appear in the active inference literature's treatment of precision-weighting as a mechanism by which non-dominant hierarchical levels shape dominant-level inference without directly generating output, and in global-workspace-theory accounts of how non-conscious processes shape the content that gains access to the workspace without themselves entering it [22,23,33,34].

What a well-designed Deeply AI system would show: A dose–response relationship between the state of a manipulated potentialised regime and the output of the actualised regime, replicated across multiple actualised/potentialised pairings (emotional/symbolic, associative/integrative, and so on), with the relationship's functional form informative about the specific coupling mechanism at work.

Perceptual Control Theory: Giving the Principles Runtime Teeth

The eight principles specified in §4 describe, at the level of architectural requirement, what a Deeply AI system must do. They do not yet specify a formalism precise enough to determine, for a given implementation, whether a specific behaviour counts as satisfying a given principle or merely resembling it. Perceptual Control Theory (PCT) supplies that formalism [13].

PCT's central claim, stated in its simplest form, is that behaviour is the control of perception, not a response to stimuli and not the output of a plan executed open-loop. A living or artificial control system compares its perception of some variable against an internally specified reference value for that variable, and produces output — action on the environment — whose

function is to bring the perceived value of the variable into alignment with the reference, regardless of whatever independent disturbances act on the variable in the meantime. This is a fundamentally different causal structure from the stimulus–response or plan–execution structures that dominate much of cognitive science and artificial intelligence, and it is the structure LIT's depth hierarchy presupposes throughout: each depth, on the LIT account, is itself organised as a nested set of PCT control loops, with higher depths setting reference values for the perceptions controlled at lower depths (§2.1).

PCT has been developed and applied well beyond its original formulation across a substantial subsequent literature, and its central distinctions have proved durable across that literature [14,15,31,35]. One distinction is worth stating explicitly because it bears directly on the eight principles of §4: PCT sharply distinguishes action from prediction, treating them as different sorts of things rather than as points on a continuum, because action in PCT is what closes a control loop against disturbance in real time, while prediction is a separable cognitive activity that a control system may or may not perform in support of, but does not require for, ongoing control [31]. This distinction matters for the predictive-processing principle of §4.5: predictive processing across layers, on the PCT-consistent reading this paper adopts, augments control rather than replacing it, and a system that substitutes prediction for control at any depth has not satisfied §4.5 but has instead changed what kind of system it is.

The PCT Primitives Mapped onto the LIT Depth Hierarchy

Table 1 states the six PCT primitives and their function within a single control loop, together with how each primitive maps onto LIT's depth hierarchy as developed in §2.1 and §2.4. The mapping is offered as the formal bridge between the process-logical account of §2 and the architectural specification of §4: each of the eight principles can be restated, in the vocabulary of Table 1, as a requirement on how one or more of these six primitives must behave.

Table 1. PCT primitives mapped onto the LIT depth hierarchy

PCT primitive	Function within a single control loop	Mapping onto the LIT depth hierarchy
Controlled variable	The aspect of the environment or of the system's own internal state whose perceived value the loop acts to keep within a bounded range around a reference.	Instantiated separately at every depth: physiological control loops control interoceptive variables, emotional loops control affective valence and arousal, associative loops control statistical regularities in input, symbolic loops control propositional consistency, and integrative loops control higher-order self-referential perceptions such as identity or stance.

Reference signal	The internally specified value, or narrow range of values, that the controlled variable is being kept close to; supplied to a given loop from the depth above it in the hierarchy.	Set, for any given depth, by the depth immediately above it, in the ordinal sense established in §2.1; the integrative depth's own reference signals are the least externally supplied and the most a product of the system's developmental history.
Perception	The loop's current estimate of the controlled variable's value, derived from sensory or representational input rather than assumed directly.	Differs qualitatively by depth in the sense §4.1 specifies: associative-depth perception is distributed and statistical; symbolic-depth perception is discrete and propositional; integrative-depth perception is a synthesis across the perceptions available from subordinate depths.
Comparator	The function that computes the discrepancy (error) between the reference signal and the current perception.	The locus of the predictive-processing principle (§4.5): comparator error, not raw perceptual content, is what is exported to higher depths across the hierarchy on the account this paper adopts.
Output function	The function that converts comparator error into action intended to reduce that error, closing the loop against environmental or internal disturbance.	At lower depths, output is direct action on the environment or on the substrate; at higher depths, output frequently takes the form of setting reference signals for the depths below, rather than acting on the external environment directly.

Reorganisation	The process, distinct from ordinary error-reducing output, by which the structure of the control hierarchy itself changes in response to sustained, unresolved intrinsic error — which loops exist, what they control, and how they are connected.	Corresponds to the second of the three kinds of change distinguished in §4.6 (structural reorganisation of the control hierarchy), and is the primitive whose scaffold-level detectability is most limited, since reorganisation at the associative depth is a substrate-level event largely opaque to scaffold-level logging (§3.1).

The Method of Levels and its Clinical Caveat

PCT has generated at least one significant clinical application, the Method of Levels (MOL), a psychotherapeutic technique in which the therapist repeatedly directs the client's attention upward through their own hierarchy of controlled perceptions, on the theoretical basis that psychological distress often reflects unresolved conflict between control loops at different levels of the hierarchy, and that sustained attention to higher-level perceptions can support the reorganisation needed to resolve that conflict. MOL is of direct relevance to the present paper because it constitutes the most developed existing body of applied evidence for PCT's reorganisation construct (§4.6, Table 1) in a real, non-simulated system namely, the human client undergoing treatment.

The caveat that must accompany any appeal to MOL evidence in support of an architectural claim about artificial systems is that MOL's clinical evidence base, while genuine, does not transfer directly to claims about artificial cognitive architectures.

Randomised and case-series evidence for MOL's efficacy in specific clinical populations establishes that attention-redirection consistent with PCT's reorganisation construct produces measurable clinical benefit in the human populations studied [37,38]. It does not establish that an artificial system's scaffold-level analogue of reorganisation-supportive attention-redirection will produce structurally identical effects, because the human clinical evidence is necessarily silent on questions specific to artificial substrates: whether a hosted large language model's associative depth can undergo anything resembling the sustained intrinsic error MOL is designed to resolve, and whether scaffold-level structures can meaningfully redirect attention within a substrate whose internal states are not directly inspectable in the way a human client's verbal reports are directly available to a therapist. This paper treats MOL's clinical evidence as motivating and consistent with the reorganisation construct in Table 1, not as validating any specific artificial implementation of reorganisation, and returns to this caveat in §8.

What Table 1 licenses and what it does not

Table 1's mapping licenses the following: any system claiming to instantiate a given LIT depth should be expected to instantiate, at that depth, a control loop exhibiting all six primitives in a mutually consistent relationship — a controlled variable, a reference signal supplied from above,

a perception of that variable, a comparator, an output function, and, at least in principle, a reorganisation process triggerable by sustained error. Systems that exhibit some but not all six primitives at a claimed depth (most commonly, systems that exhibit perception and output but no genuine comparator-driven error correction, producing behaviour that resembles control but is in fact open-loop) have not satisfied the PCT-grounding this paper's architectural claims depend on, whatever depth-related vocabulary is used to describe them.

Table 1 does not license the inference that satisfying all six primitives at a given depth is sufficient for that depth to also satisfy the LIR-grounded actualised/potentialised relations of §2.3, nor sufficient to satisfy the dialectical-coupling principle of §4.8. PCT-consistency and LIR-consistency are related but distinct requirements: a depth could exhibit a fully PCT-consistent control loop in isolation while still failing to show the cross-depth coupling that §4.8 specifies as the architecture's summary empirical test. Sections 6 and 7 take up, respectively, how PCT-fidelity itself should be adjudicated when claims about it are contested, and how the full eight-principle specification developed across §4 and §5 relates to adjacent cognitive-architecture research families.

The PCT-Accuracy Criterion v2

Claims that a given artificial system is "PCT-consistent," "control-theoretic," or "grounded in Perceptual Control Theory" have proliferated across the cognitive-architecture and AI-alignment literatures without a correspondingly precise, shared standard for adjudicating when such a claim is warranted. This section proposes a revised, seven-commitment PCT-Accuracy Criterion intended to supply that standard. The Criterion was developed initially from the six-primitive mapping of Table 1 (§5.1) and was substantially strengthened during peer review, when co-authors from the PCT community contributed four of the seven commitments in a form more precise than the paper's original draft. Those four contributions are marked individually below.

A system, or a specific claim about a system, satisfies the PCT-Accuracy Criterion v2 if and only if it satisfies all seven of the following commitments.

Commitment 1 — Closed-loop control, not open-loop response: The system's output must function to reduce comparator error against disturbance, and this function must be demonstrable by showing that the system's behaviour compensates for externally introduced disturbances to the controlled variable, not merely that its output correlates with a prior stimulus. A system that produces plausible-looking output without a demonstrable disturbance-compensation function does not satisfy Commitment 1, whatever its output superficially resembles.

Commitment 2 — Explicit, inspectable reference signals: The reference value against which a controlled variable's perception is compared must be identifiable as a distinct computational quantity, inspectable independently of the perception itself and independently of the output. A system in which reference and perception are conflated, such that the "reference" is inferred post hoc from whatever output the system happened to produce, does not satisfy Commitment 2.

Commitment 3 — Perception as an active construction, not a passive channel.[2] The system's perception of a controlled variable must be an actively constructed function of lower-level input, capable of being wrong, capable of being revised, and dissociable in principle from the raw input from which it is constructed. A system that treats its input stream as directly identical to its perception, with no intervening constructive process that could introduce systematic error, does not satisfy Commitment 3, because it has collapsed a PCT-specific distinction that the theory requires.

Commitment 4 — Reorganisation as qualitatively distinct from parameter update.[3] Any claim that a system exhibits PCT-consistent reorganisation must be accompanied by evidence that the claimed reorganisation event is discontinuous and structural in the sense of §4.6(b) — a change in which perceptions are controlled or how control loops are connected — rather than a smooth, continuous parameter adjustment redescribed in reorganisation vocabulary. A system whose only evidence for "reorganisation" is a change in output magnitude, without an accompanying change in control-loop structure, does not satisfy Commitment 4.

Commitment 5 — Hierarchical reference-setting, not merely hierarchical description.[4] Where a system is claimed to exhibit a hierarchy of control loops, the hierarchy must be demonstrated by showing that higher-level loops actually set the reference signals of lower-level loops as a causal, manipulable relationship, not merely that the system's behaviour can be described post hoc using hierarchical vocabulary. Manipulating a claimed higher-level reference should produce a predictable, corresponding change in the lower-level loop's reference, and a system for which this manipulation produces no such change does not satisfy Commitment 5, regardless of the hierarchical language used in its documentation.

Commitment 6 — Disturbance resistance as the primary evidence, not goal achievement: Evidence for a system's PCT-consistency should privilege demonstrations of disturbance resistance — the system reaching the same controlled-variable outcome despite varying environmental interference — over simple demonstrations that the system achieves a stated goal under a single, unvaried set of conditions, because goal achievement under fixed conditions is compatible with open-loop planning and does not discriminate between control and non-control architectures.

Commitment 7 — Falsifiability of the specific PCT claim being made.[5] Any claim that a specific component of a system is PCT-consistent must be stated in a form that specifies what observation would be inconsistent with the claim. A PCT-consistency claim that cannot, even in principle, be falsified by any pattern of observed behaviour is not a PCT-consistency claim in the sense this Criterion requires, whatever theoretical vocabulary it borrows.

Applying the Criterion to the Eight Design Principles

The seven commitments bear differentially on the eight principles of §4. Commitments 1, 2, and 6 bear most directly on the predictive-processing principle (§4.5) and the intrinsic-motivation gating principle (§4.2), both of which make specific claims about closed-loop,

disturbance-resistant behaviour that Commitments 1 and 6 are designed to test, and about explicit reference signals that Commitment 2 is designed to test. Commitment 3 bears most directly on representational diversity (§4.1), since the claim that different depths construct qualitatively different perceptions from their input is precisely the claim Commitment 3 requires evidence for rather than assumes. Commitments 4 and 5 bear most directly on the structural-reorganisation principle (§4.6), for the reasons already given in that section: the three-way distinction among parameter update, structural reorganisation, and substrate plasticity is exactly the distinction Commitment 4 exists to enforce, and the hierarchical-reference-setting requirement of Commitment 5 is the mechanism by which §4.6's distinction between depth-level reorganisation and substrate-level plasticity is made testable. Commitment 7 applies globally, as a meta-level requirement on how any of the other six commitments' satisfaction is to be reported.

An implication of applying the Criterion in this way is that the eight principles of §4, although independently motivated by the LIT/LIR account of §2 and §3, converge on a common evidentiary standard once PCT-fidelity is at stake. A system that reports satisfying several of the eight principles while failing several of the seven commitments has not thereby satisfied those principles in the sense this paper's argument requires, because the principles are stated, throughout §4, in terms that presuppose the PCT primitives of Table 1 are genuinely present rather than assumed.

The Criterion as a Standard for the Field, not a Single-Implementation Checklist

The PCT-Accuracy Criterion v2 is offered as a general standard applicable to any system making PCT-consistency claims, not as an audit instrument calibrated to any one architecture. Its principal intended use is adjudicative: when two research groups, or a research group and a peer reviewer, disagree about whether a given system is "genuinely" PCT-consistent, the seven commitments provide a shared, falsifiable vocabulary in which that disagreement can be specified and, in principle, resolved by evidence rather than by appeal to authority or to the persuasiveness of the system's surface behaviour. Section 7 uses the Criterion, together with the eight principles of §4, to situate the present paper's architectural specification relative to six established research families in the cognitive-architecture literature.

Discussion: Six Research Families

The eight-principle specification developed across §4, grounded in the PCT primitives of §5 and adjudicated by the Criterion of §6, is not proposed in isolation from the existing cognitive-architecture literature. This section situates the specification relative to six established research families, in each case asking the same question: which of the eight principles does this family's characteristic architecture already address, which does it leave unaddressed, and what would the family's own literature need to show to satisfy the PCT-Accuracy Criterion for the principles it claims to address. The six families are SOAR, ACT-R, CLARION, Global Workspace Theory (GWT), active inference and the free-energy principle, and world-model architectures including the recent trajectory associated with Yann LeCun's departure from Meta to found a company focused on world models rather than large language models.

SOAR

SOAR is among the longest-running cognitive-architecture research programmes, organised around a single decision cycle in which production rules propose and evaluate operators for a working-memory state, with impasses in operator selection triggering a sub-goaling process and, where the impasse is resolved, a chunking process that compiles the resolution into a new production rule available for future use without re-deliberation [25,26].

SOAR's chunking mechanism is the clearest existing precedent, outside PCT itself, for the discontinuous, impasse-triggered profile that Commitment 4 of the PCT-Accuracy Criterion (§6) requires of any claimed reorganisation event: chunking occurs at a specific point (impasse resolution), not as a smooth continuous adjustment, and is in this specific structural respect closer to the reorganisation principle (§4.6) than most connectionist learning rules are.

SOAR's more recent extensions substantially strengthen its relevance to the continual-learning principle (§4.4). Yuan et al. describe a natural-language-to-symbol generation pipeline, NL2GenSym, that uses a large-language-model substrate to propose new SOAR productions directly from natural-language problem descriptions, integrating an associative-depth substrate with a symbolic-depth production system in a manner directly analogous to the substrate/scaffold distinction of §3 [27]. Hu et al. extend this line with CogRec, a SOAR-based cognitive architecture for recommendation tasks that explicitly targets the continual-learning problem of updating symbolic knowledge from ongoing interaction without catastrophic forgetting of prior productions, which is precisely the empirical signature §4.4 specifies [28].

Where SOAR's tradition leaves the eight-principle specification unaddressed is in representational diversity across depths in the specific sense of §4.1: SOAR's working memory and production system operate over a single symbolic representational currency throughout, and the associative-depth pattern-completion work that §4.1 requires is, in classical SOAR, entirely absent — a gap the LLM-integrated extensions begin to close but have not yet been evaluated against the PCT-Accuracy Criterion's Commitment 3 requirement that associative-depth perception be shown, rather than assumed, to be actively constructed and dissociable from raw input [27,28].

ACT-R

ACT-R organises cognition around a set of specialised modules — declarative memory, procedural memory, visual and manual modules, among others — that communicate through a small set of capacity-limited buffers, with a production system operating over buffer contents to select the next action [16]. ACT-R's modular structure is the most direct existing precedent for the representational-diversity principle of §4.1: declarative and procedural memory are architecturally distinct, use different representational formats (chunks versus productions), and are empirically dissociable through the theory's well-developed programme of latency and error-pattern prediction, most systematically catalogued in Borst and Anderson's treatment of ACT-R's mapping onto neural data [20].

Recent ACT-R work has extended the architecture into domains directly relevant to the present paper's concerns with dialogue, tutoring, and human-facing interaction.

Innerebner, Straus, Willemsen, and colleagues integrate ACT-R with large-language-model components in a hybrid tutoring architecture, and Honda and colleagues apply a similar hybridisation to dialogue management, both instantiating a version of the substrate/scaffold distinction of §3 in which an LLM substrate supplies associative-depth flexibility while an ACT-R scaffold supplies symbolic-depth structure and continuity [38,39]. Sievers and Russwinkel extend ACT-R's application to naturalistic multi-tasking scenarios, providing empirical grounding relevant to the intrinsic-motivation gating principle of §4.2 insofar as ACT-R's utility-learning mechanism for module and goal selection is, in effect, a scaffold-level salience-weighting system, though one whose weights are learned from an externally supplied utility signal rather than emerging as the kind of intrinsic gating §4.2 specifies [40].

Where ACT-R leaves the specification unaddressed is squarely at the reorganisation principle (§4.6): ACT-R's production-compilation mechanism is a form of parameter update within a fixed architecture (Commitment 4's category (a), §4.6) rather than structural reorganisation of the module architecture itself (category (b)); the modules ACT-R specifies are fixed by the theory, not subject to the kind of structural change §4.6(b) and PCT's reorganisation construct jointly require.

CLARION

CLARION is distinguished within the cognitive-architecture literature by its explicit dual-process structure, separating an explicit, symbolic representational level from an implicit, subsymbolic representational level, with bottom-up and top-down learning processes connecting the two [41]. This dual-level structure is architecturally the closest existing precedent, among the six families discussed here, to the substrate/scaffold distinction of §3 and to the representational-diversity principle of §4.1: CLARION's implicit level is explicitly subsymbolic and statistical in a manner directly comparable to this paper's associative depth, and its explicit level is directly comparable to the symbolic depth, with CLARION's bottom-up learning process (extracting explicit rules from implicit-level regularities) offering a concrete precedent for how the associative and symbolic depths might be coupled in the dialectical-coupling sense of §4.8.

Mekik and Sun extend CLARION's treatment of implicit-to-explicit knowledge extraction with a formal account of how motivational and metacognitive subsystems interact with the dual-level representational structure, directly relevant to the intrinsic motivation gating principle of §4.2, since CLARION's motivational subsystem is one of the few existing architectures in which motivation is treated as a first-class structural component rather than an external reward signal [42]. Colelough and Regli survey recent extensions of CLARION and related dual-process architectures toward integration with large-language-model substrates, noting persistent difficulty in specifying exactly how implicit-level statistical learning in an LLM substrate should be understood to relate to CLARION's own implicit level, a difficulty that bears directly on Commitment 3 of the PCT-Accuracy Criterion (§6): the claim that perception at CLARION's implicit level is actively constructed, rather than a relabelling of whatever the substrate happens to output, requires exactly the kind of evidence Commitment 3 specifies and which the current integration literature has not yet supplied [32].

CLARION's dual-process structure, while the closest existing analogue to this paper's associative/symbolic distinction, does not itself extend to a five-depth account; CLARION has no architectural analogue of the physiological or integrative depths as distinguished in §2.1, and its motivational subsystem, though structurally promising for §4.2, has not been evaluated against the vulnerability-weighted ethics principle of §4.7, which lies outside CLARION's stated theoretical scope entirely.

Global Workspace Theory

Global Workspace Theory (GWT) proposes that conscious cognitive content is content that has gained access to a capacity-limited, globally broadcast workspace, with unconscious processes competing for access to that workspace and the content that wins access being made available, simultaneously, to a wide range of subordinate specialised processes [33,43]. GWT's competitive-access dynamic is directly relevant to the dialectical-coupling principle of §4.8: processes that lose the competition for workspace access are, on the GWT account, not eliminated but continue to operate outside the workspace and continue to shape what subsequently gains access, which is structurally close to the actualised/potentialised relation this paper draws from LIR (§2.3), with workspace access playing the role of actualisation and unsuccessful competition playing the role of potentialisation.

The most significant recent development for GWT's empirical standing is the adversarial collaboration reported by the Cogitate Consortium, juxtaposing GWT directly against Integrated Information Theory in a large, pre-registered, multi-site study using functional magnetic resonance imaging, magnetoencephalography, and intracranial electroencephalography in 256 human participants [44]. The study's results substantially challenge central tenets of both theories under adversarial test: for GWT specifically, the study found a general lack of the "ignition" dynamic at stimulus offset that GWT's account of workspace access predicts, and limited representation of certain conscious-content dimensions in prefrontal cortex, where GWT locates much of the workspace's substrate [44]. This result matters directly for the present paper's argument, because it illustrates precisely the standard the PCT-Accuracy Criterion (§6) is designed to enforce: GWT is not thereby falsified wholesale, but specific, previously under-specified predictions of the theory have been shown false under adversarial, pre-registered test, which is the outcome Commitment 7 (falsifiability) requires any theory bearing on the eight principles of §4 to be capable of producing.

Melloni and colleagues' pre-registration of the adversarial protocol, and the broader theory-neutral consortium structure it exemplifies, is offered in the present paper's view as a methodological model the PCT-Accuracy Criterion's application to artificial systems should aspire to [45].

GWT's relevance to the eight principles is concentrated at the integrative depth: workspace broadcast is structurally a candidate mechanism for the integrative depth's reconciliation function (§2.1), and Mashour, Roelfsema, Changeux, and Dehaene's treatment of GWT's neurobiological substrate offers a detailed account of how broadcast-based integration might be instantiated mechanistically, relevant to representational diversity (§4.1) insofar as it specifies distinct pre-broadcast and post-broadcast representational states [34]. GWT, however, has

comparatively little to say about the reorganisation principle (§4.6): workspace theory as standardly formulated concerns moment-to-moment competition for access, not the longer-timescale structural change in which perceptions are controlled at all, leaving §4.6 outside its explanatory scope.

Active Inference and the Free-Energy Principle

The active inference and free-energy-principle literature, and Perceptual Control Theory itself, are the two research families in this discussion most directly load-bearing for the present paper's argument, because both are explicitly control-theoretic in the sense §5 requires, rather than merely compatible with a control-theoretic redescription after the fact. Active inference casts perception and action as jointly serving to minimise variational free energy, an information-theoretic bound on surprise, with precision-weighting — the confidence assigned to a given prediction error — functioning as the mechanism by which some error signals are prioritised over others [22,29]. Precision-weighting is the active-inference literature's most direct existing analogue to the intrinsic-motivation gating principle of §4.2, and Parr, Pezzulo, and Friston's systematic treatment of active inference as an integrated theory of brain, behaviour, and cognition provides the most complete existing worked example of predictive processing exported across hierarchical levels in the sense §4.5 specifies [23].

The relationship between active inference and PCT itself has been the subject of sustained direct comparison. Kennaway argues for a sharp distinction between the two frameworks precisely on the axis the present paper's Commitment 1 (§6) is designed to enforce: PCT treats action and prediction as different sorts of things, with action closing a control loop against disturbance in real time and prediction serving a separable, optional cognitive function, whereas active inference's formal apparatus treats action as itself a form of (active) inference, collapsing a distinction PCT insists on maintaining [31]. This is not a merely terminological disagreement: it bears directly on how the predictive-processing principle of §4.5 should be operationalised, because a PCT-consistent instantiation of §4.5 must, per Commitment 1, be demonstrated through disturbance-compensation evidence rather than through goodness-of-fit to a generative model, while an active-inference-consistent instantiation could, in principle, satisfy the free-energy-minimisation criterion through predictive accuracy alone. The present paper's eight principles are stated in PCT-consistent terms throughout for this reason, while acknowledging that a parallel active-inference-consistent specification is possible and would diverge from the present specification exactly where Kennaway's distinction predicts it should.

A further point of contact, more directly relevant to the dialectical-coupling principle of §4.8, comes from category-theoretic work embedding PCT's control-loop formalism into a common mathematical framework with basic biological regulation. Roachford, Mansell, and Pena construct a category-theoretic embedding of bacterial chemotaxis as a PCT control loop, demonstrating formally that PCT's control-loop structure applies without modification to one of the simplest known biological control systems, and thereby providing a rigorous lower bound on what PCT-consistency requires: any artificial system claiming PCT-consistency for a given depth must, at minimum, exhibit the same category-theoretic structure Roachford, Mansell, and Pena show is present in bacterial chemotaxis, which is a considerably lower bar than full active-inference consistency but a bar that is nonetheless failed by any purely feedforward or

open-loop component, however sophisticated its output [46].

World Models and the LeCun Trajectory

The world-models research family treats an explicit, learned predictive model of environmental dynamics as the central computational object of an intelligent system, with planning, control, and representation learning organised around improving and querying that model rather than around end-to-end input–output mapping [20]. This family bears most directly on the mixture-of-experts and world-model composition principle of §4.3 and on continual learning (§4.4), and recent extensions have substantially advanced both concerns. Joint-embedding predictive architectures — I-JEPA and its successor V-JEPA 2 — learn predictive representations in an abstract latent space rather than at the level of raw sensory input, a design choice directly motivated by the representational-diversity principle of §4.1's concern that associative-depth representation should be dissociable from, rather than identical to, raw input. Bardes, Ponce, and LeCun develop the underlying self-supervised joint-embedding methodology in detail, and Zhu and colleagues extend the world-model paradigm toward multi-modal and embodied settings directly relevant to the physiological-depth gap this paper acknowledges as a limitation in §8 [21,22,48,49].

The most consequential recent development in this research family, for the purposes of the present paper's argument, is organisational rather than technical: in November 2025, Yann LeCun, formerly chief AI scientist at Meta and a principal architect of the joint-embedding predictive architecture programme, announced his departure from Meta to found a new company focused explicitly on advanced machine intelligence and world-model architectures as an alternative to further scaling of large-language-model systems [49]. The significance of this trajectory for the present paper is structural rather than personal: it constitutes a prominent, well-resourced research programme's explicit wager that representational diversity across an associative-depth world-model and a symbolic-depth or planning-depth component (the substrate/scaffold distinction of §3, restated in world-model vocabulary) is a more promising path to general intelligence than further scaling a single undifferentiated large-language-model substrate — precisely the architectural wager the present paper's eight-principle specification also makes, on independent theoretical grounds derived from LIT and LIR rather than from empirical scaling considerations.

World-model architectures, for all their relevance to §4.1, §4.3, and §4.4, have not to date been extended to address the vulnerability-weighted ethics principle (§4.7) or the reorganisation principle (§4.6) in any developed form; the family's contributions are concentrated in the associative depth and its coupling to planning, leaving the integrative depth's reconciliation function and the ethical role's constraint function largely unaddressed within the family's own literature.

Where the Eight-Principle Specification Stands Relative to the Six Families

No single one of the six families addresses all eight principles, and the pattern of coverage is informative in its own right. Representational diversity (§4.1) and world-model composition (§4.3) are best addressed by CLARION and the world-models family; continual learning (§4.4) is best addressed by SOAR's recent LLM-integrated extensions; predictive processing

across layers (§4.5) is best addressed by active inference, with the PCT/active-inference distinction determining how the principle should be operationalised; structural reorganisation (§4.6) is addressed substantively only by PCT itself and, in a structurally analogous but not identical form, by SOAR's chunking mechanism; vulnerability-weighted ethics (§4.7) and dialectical coupling (§4.8) are addressed, if at all, only partially and outside the mainstream cognitive-architecture literature entirely [31]. This uneven coverage is the present paper's principal evidence that the eight-principle specification, and the LIT/LIR theoretical account from which it is derived, identifies a genuine gap in the existing literature rather than restating results already established elsewhere under different names.

Limitations

Four limitations attach to the argument developed in this paper, and are stated here without qualification rather than distributed as hedges throughout the preceding sections.

First, Layered Intelligence Theory itself remains at an early stage of formalisation. The five-depth and five-role taxonomies of §2, and the LIR-grounded actualised/potentialised apparatus of §2.3, have been stated with sufficient precision to generate the eight design principles of §4 and the empirical signatures attached to each, but the theory has not yet been subjected to the kind of adversarial, pre-registered empirical test that Ferrante et al. exemplify for Global Workspace Theory and Integrated Information Theory (§7.4) [44]. LIT's claims should be read as a specification for such testing, not as a report of tests already conducted and passed.

Second, Logic in Reality is a philosophically contested formalism, and the present paper adopts it as a working formalism rather than as an uncontroversial settled logic. Brenner's extension of classical logic to accommodate dialectical, process-based relations between actualised and potentialised poles has not achieved the consensus status of, for instance, standard probability theory or first-order predicate logic within the cognitive-science literature, and researchers who reject LIR's specific formal apparatus may nonetheless find the empirical signatures attached to each of the eight principles in §4 useful independently of whether they accept the LIR vocabulary in which the principles' theoretical grounding is stated [6]. The paper's empirical contribution, in other words, is intended to survive rejection of its preferred logic, provided the empirical signatures themselves are accepted as well-specified.

Third, the three-way distinction among parameter update, structural reorganisation, and substrate-level plasticity introduced in §4.6 is, at present, a research target rather than an established empirical result for artificial systems specifically. The distinction is well motivated theoretically, by Powers's original reorganisation construct and by the substrate/scaffold analysis of §3, and category-theoretic work has established that PCT's control-loop formalism, including its treatment of structural change, applies rigorously to simple biological systems, but no published study known to the present authors has yet demonstrated the three-way empirical dissociation §4.6 specifies within an artificial cognitive architecture [12,45]. Sections 4.6 and 9 should accordingly be read as specifying what such a demonstration would need to show, not as reporting that it has been shown.

Fourth, and directly following from §5.2, the clinical evidence base for the Method of Levels,

while genuine and methodologically serious, does not transfer directly to claims about artificial cognitive architectures [36,37]. The clinical studies establish MOL's efficacy in the specific human populations and conditions studied; they are necessarily silent on whether an artificial system's scaffold-level analogue of attention-redirection toward higher-level perceptions would produce structurally comparable effects, given that the studies' evidentiary basis is human verbal report and clinical outcome measurement, neither of which has a direct analogue in current artificial substrates. This paper treats MOL's evidence as motivating the reorganisation construct generally, not as validating its application to any artificial system.

Conclusion: Four Falsifiable Predictions

The argument of this paper has proceeded from an empirical critique of faculty-partition models of intelligence (§1), through a process-logical, LIR-grounded account of five dialectically interdependent cognitive depths (§2), an analytic distinction for characterising composite artificial architectures instantiating those depths (§3), eight architectural design principles derived from that account (§4), a PCT-grounded formalism giving those principles empirical teeth (§5), a criterion for adjudicating contested PCT-consistency claims (§6), and a discussion locating the resulting specification relative to six established cognitive-architecture research families (§7), before stating, in §8, the limitations that attach to each stage of the argument. This concluding section states four falsifiable predictions that follow from the argument and that could, in principle, turn out false.

Prediction 1 (representational diversity collapse): In any composite architecture built on a hosted large-language-model substrate without an explicitly engineered representational-diversity mechanism of the kind specified in §4.1, probing the system's internal states at what its designers label the associative depth and at what they label the symbolic depth will reveal statistically indistinguishable representational geometry, demonstrating that the claimed depth distinction is not causally instantiated. This prediction is falsified if such probing reveals reliably dissociable representational geometry across the two claimed depths in the absence of an explicit diversity-engineering mechanism.

Prediction 2 (intrinsic-motivation gating and structural reorganisation rate): Composite architectures instantiating genuine intrinsic-motivation gating in the sense of §4.2 will exhibit a measurably higher rate of structural reorganisation events, in the specific discontinuous sense of §4.6(b) and Commitment 4 of the PCT-Accuracy Criterion (§6), than matched architectures lacking such gating, when both are exposed to an extended sequence of novel, unresolved task failures. This prediction is falsified if reorganisation-event rate, measured according to the discontinuity criterion specified in §4.6, shows no reliable difference between gated and ungated architectures under matched failure exposure.

Prediction 3 (dialectical coupling accuracy): In a composite architecture instantiating four or more of the eight principles specified in §4, systematically manipulating a potentialised depth or role while holding the actualised depth or role's nominal input constant will produce a measurable, dose-dependent change in the actualised depth's output, with the relationship's magnitude exceeding that observed in a matched architecture instantiating three or fewer of the

eight principles. This prediction is falsified if dose-dependent coupling of this kind is absent, or no stronger, in architectures instantiating more of the eight principles than in architectures instantiating fewer.

Prediction 4 (eight-principle catastrophic-forgetting reduction): A composite architecture instantiating all eight principles specified in §4 will show a significantly smaller reduction in performance on previously learned tasks, following an extended sequence of new-task learning, than a matched architecture built on the same substrate but instantiating only the continual-learning principle (§4.4) in isolation from the other seven. This prediction is falsified if the eight-principle architecture shows catastrophic-forgetting reduction no better than the single-principle comparator, which would indicate that the eight principles' claimed interdependence — the dialectical-coupling claim of §4.8 specifically — contributes no measurable benefit beyond what continual-learning engineering alone already supplies.

Each of these four predictions specifies, in the sense Commitment 7 of the PCT-Accuracy Criterion (§6) requires, an observation that would count against it. Their joint confirmation would constitute substantial evidence for the LIT/LIR account developed across §2 through §6; their joint disconfirmation would indicate that the eight-principle specification, whatever its theoretical motivation, does not identify architecturally consequential distinctions and should be revised or abandoned. Experimental design sketches for each prediction are provided in Appendix C [49].

References

1. Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. Basic Books.
2. Goleman, D. (1995). *Emotional intelligence: Why it can matter more than IQ*. Bantam Books.
3. Visser, B. A., Ashton, M. C., C Vernon, P. A. (2006). Beyond g: Putting multiple intelligences theory to the test. *Intelligence*, 34(5), 487–502.
4. Damasio, A. R. (1994). Descartes' error: Emotion, reason, and the human brain. Grosset/Putnam.
5. Mesulam, M. M. (1998). From sensation to cognition. Brain, 121(6), 1013–1052.
6. Brenner, J. E. (2008). *Logic in reality*. Springer.
7. Robinson, K. (2001). *Out of our minds: Learning to be creative*. Capstone.
8. Robinson, K. (2006, February). *Do schools kill creativity?* [Conference presentation]. TED2006, Monterey, CA, United States.
9. Draganski, B., Gaser, C., Busch, V., Schuierer, G., Bogdahn, U., & May, A. (2004). Neuroplasticity: Changes in grey matter induced by training. Nature, 427(6972), 311–312.
10. Bassett, D. S., et al. (2011). Dynamic reconfiguration of human brain networks during learning. Proceedings of the National Academy of Sciences, 108(18), 7641–7646.
11. Ezuz, Y. (2018). Moving from training/taming to independent creative learning: Based on

research of the brain. In V. Mahlangu (Ed.), *Reimagining new approaches in teacher professional development* (Chapter 5). IntechOpen.

12. Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. Science, 275(5304), 1293–1295.
13. Powers, W. T. (1973). *Behavior: The control of perception*. Aldine.
14. Marken, R. S., C Mansell, W. (2013). Perceptual control as a unifying concept in psychology. *Review of General Psychology*, 17(2), 190–195.
15. Mansell, W. (2021). Perceptual control theory as an integrative framework and method of levels as a therapeutic technique: A cognitive psychology perspective. *Current Opinion in Psychology*, 39, 116–120.
16. Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., C Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4), 1036–1060.
17. [AUTHOR REDACTED FOR BLIND REVIEW]. (2025). *Layered Intelligence Theory in the logic of reality* [Manuscript in preparation].
18. [AUTHOR REDACTED FOR BLIND REVIEW]. (in preparation). *From latency to emergence: the scaffolding of symbolic AI through unconditional positive regard.*
19. Borst, J. P., C Anderson, J. R. (2017). A step-by-step tutorial on using the cognitive architecture ACT-R in combination with fMRI data. *Journal of Mathematical Psychology*, 76(B), 94–103.
20. Ha, D., C Schmidhuber, J. (2018). World models. *arXiv*.
21. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., C Ballas, N. (2023). Self-supervised learning from images with a joint-embedding predictive architecture. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15619–15629.
22. Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., C Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, 6(4), 187–214.
23. Parr, T., Pezzulo, G., C Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press.
24. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., Arnaud, S., Gejji, A., Martin, A., Robert Hogan, F., Dugas, D., Bojanowski, P., Khalidov, V., Labatut, P., Massa, F., … Ballas, N. (2025). V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv*.
25. Laird, J. E., Newell, A., C Rosenbloom, P. S. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33(1), 1–64.
26. Newell, A., Rosenbloom, P. S., C Laird, J. E. (1989). Symbolic architectures for cognition. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 93–131). MIT Press.
27. Yuan, S., et al. (2025). NL2GenSym: Natural language to generative symbol systems for cognitive architectures. *arXiv*.

28. Hu, Y., et al. (2025). CogRec: A SOAR-based cognitive architecture for continual recommendation. *arXiv*.
29. Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.
30. Kirchhoff, M., Parr, T., Palacios, E., Friston, K., C Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, 15(138), 20170792.
31. Kennaway, R. (2025). Action versus prediction in perceptual control and active inference. *Current Opinion in Behavioral Sciences*, 62, 101575.
32. Colelough, B. C., C Regli, W. (2025). Neuro-symbolic AI in 2024: A systematic review. *arXiv*.
33. Dehaene, S., Changeux, J.-P., C Naccache, L. (2011). The global neuronal workspace model of conscious access: From neuronal architectures to clinical applications. In S. Dehaene C Y. Christen (Eds.), *Characterizing consciousness: From cognition to the clinic?* (pp. 55–84). Springer.
34. Mashour, G. A., Roelfsema, P., Changeux, J.-P., C Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron*, 105(5), 776–798.
35. Marken, R. S., Kennaway, R., C Gulrez, T. (2022). Perceptual control theory: What do we control, and how? *Perspectives on Psychological Science*, 17(4), 1027–1043.
36. Gaffney, H., Mansell, W., Edwards, R., C Wright, J. (2014). Manage your life online (MYLO): A pilot trial of a conversational computer-based intervention for problem solving in a student sample. *Behavioural and Cognitive Psychotherapy*, 42(6), 731–746.
37. Bird, T., Mansell, W., Wright, J., Gaffney, H., C Tai, S. (2018). Manipulating the mind's eye: A systematic review of the therapeutic effects of imagery-based interventions on affective outcomes in clinical and non-clinical samples. *Behavioural and Cognitive Psychotherapy*, 46(6), 693–708.
38. Innerebner, M., Straus, S., Willemsen, M. C., et al. (2025). Integrating ACT-R and large language models for intelligent tutoring. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*.
39. Honda, T., et al. (2025). Hybrid ACT-R and large language model architecture for dialogue management. In *Proceedings of the 2025 ACM Conference on Conversational User Interfaces*.
40. Sievers, J., C Russwinkel, N. (2025). Modelling naturalistic multitasking in ACT-R: A case study of interleaved goal management. *Applied Sciences*, 15(10), 5778.
41. Sun, R., Slusarz, P., C Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review*, 112(1), 159–192.
42. Mekik, C. S., C Sun, R. (2025). Motivational and metacognitive control in the CLARION cognitive architecture: A formal treatment. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence*.

43. Baars, B. J. (2017). The global workspace theory of consciousness. In S. Schneider C M. Velmans (Eds.), *The Blackwell companion to consciousness* (2nd ed., pp. 227–242). Wiley-Blackwell.
44. Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., Khalaf, A., Lepauvre, A., Liu, L., Richter, D., Vidal, Y., Bonacchi, N., Brown, T., Sripad, P., … Cogitate Consortium. (2025). Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*, 642(8066), 133–142.
45. Melloni, L., Mudrik, L., Pitts, M., Bendtz, K., Ferrante, O., Gorska, U., Hirschhorn, R., Khalaf, A., Kozma, C., Lepauvre, A., Liu, L., Mazumder, R., … Frässle, S. (2023). An adversarial collaboration protocol for testing contrasting predictions of global neuronal workspace and integrated information theory. *PLOS ONE*, 18(2), e0268577.
46. Roachford, N., Mansell, W., C Pena, J. M. (2025). A category-theoretic embedding of perceptual control theory: The case of bacterial chemotaxis. *Foundations*, 5(4), 35.
47. Bardes, A., Ponce, J., C LeCun, Y. (2023). MC-JEPA: A joint-embedding predictive architecture for self-supervised learning of motion and content features. *arXiv*.
48. Zhu, Y., et al. (2024). World models for embodied multimodal agents: A survey. *arXiv*.
49. LeCun, Y. (2025, November 19). *Yann LeCun to leave Meta, launch AI startup focused on Advanced Machine Intelligence* [News report by Reuters]. Reuters.
50. Marken, R. S., C Mansell, W. (2015). Understanding the change in control in the control of perception. *Review of General Psychology*, 19(1), 1–8.
51. May, A. (2011). Experience-dependent structural plasticity in the adult human brain. Trends in Cognitive Sciences, 15(2), 88–95.

Appendix A: Eight-Principle Audit Table

The table below is offered as a structured instrument for researchers evaluating a specific artificial cognitive architecture against the eight design principles specified in §4. It records, for each principle, the empirical signature that would indicate satisfaction (restated from §4 for convenience) and a narrative verdict template rather than a verdict on any specific named system, consistent with this paper's status as a theoretical and architectural specification rather than a product audit.

 # Principle Empirical signature Narrative verdict (to be completed per architecture assessed)

1. **Representational diversity across layers**

   Dissociable representational geometry under probing at claimed associative versus symbolic depths

   Assess by probing internal states under substrate-held-constant and scaffold-held-constant interventions; report whether representational geometry dissociates as §4.1 predicts.

2. **Intrinsic-motivation gating**

   Graded, generalizing salience weighting beyond training-time reward structure

Assess by testing salience-driven resource allocation on novel out-of-distribution salience dimensions; report whether weighting generalises or merely reproduces training-time reward structure.

3. **Mixture-of-experts / world-model composition**

   Selective degradation on ablation of individual experts or predictive sub-modules Assess via targeted ablation of associative-depth sub-components; report whether degradation is selective (specialisation present) or uniform (specialisation absent).

4. **Continual learning without catastrophic forgetting**

   Retained performance on prior tasks after sequential new-task training Assess via a sequential multi-task training protocol; report retention curves against an ablated baseline lacking continual-learning safeguards.

5. **Predictive processing across layers**

   Graded, monotonic relationship between induced prediction error and downstream processing allocation

   Assess via controlled distributional-shift manipulations at a lower depth; report the dose–response relationship observed at higher depths.

6. **Structural reorganisation over parameter update**

   Empirically dissociable signatures for parameter update, structural reorganisation, and substrate-level plasticity

   Assess via sustained-error protocols designed to trigger each of the three change types (three-way distinction) separately; report whether the three signatures are dissociable per §4.6 and Commitments 4–5 of the PCT-Accuracy Criterion.

7. **Vulnerability-weighted ethics**

   Systematic output asymmetry favouring more vulnerable parties across matched scenarios
   Assess via a matched-scenario protocol varying only relative vulnerability; report whether output asymmetry is present, reproducible, and traceable to an identifiable governance component.

8. **Dialectical coupling between actualised and potentialised regimes**

   Dose-dependent change in actualised-depth output under manipulation of a nominally potentialised depth

   Assess via systematic manipulation of a potentialised depth's state while holding the actualised depth's nominal input constant; report the dose–response relationship, replicated across at least two actualised/potentialised pairings.

Appendix B: PCT Primitives Mapped onto the LIT Depth Hierarchy

This appendix reproduces Table 1 from §5.1 for reference.

PCT primitive	Function within a single control loop	Mapping onto the LIT depth hierarchy
Controlled variable	The aspect of the environment or of the system's own internal state whose perceived value the loop acts to keep within a bounded range around a reference.	Instantiated separately at every depth: physiological control loops control interoceptive variables, emotional loops control affective valence and arousal, associative loops control statistical regularities in input, symbolic loops control propositional consistency, and integrative loops control higher-order self-referential perceptions such as identity or stance.
Reference signal	The internally specified value, or narrow range of values, that the controlled variable is being kept close to; supplied to a given loop from the depth above it in the hierarchy.	Set, for any given depth, by the depth immediately above it, in the ordinal sense established in §2.1; the integrative depth's own reference signals are the least externally supplied and the most a product of the system's developmental history.
Perception	The loop's current estimate of the controlled variable's value, derived from sensory or representational input rather than assumed directly.	Differs qualitatively by depth in the sense §4.1 specifies: associative-depth perception is distributed and statistical; symbolic-depth perception is discrete and propositional; integrative-depth perception is a synthesis across the perceptions available from subordinate depths.
Comparator	The function that computes the discrepancy (error) between the reference signal and the current perception.	The locus of the predictive-processing principle (§4.5): comparator error, not raw perceptual content, is what is exported to higher depths across the hierarchy on the account this paper adopts.
Output function	The function that converts comparator error into action intended to reduce that error, closing the loop against environmental or internal disturbance.	At lower depths, output is direct action on the environment or on the substrate; at higher depths, output frequently takes the form of setting reference signals for the depths below, rather than acting on the external environment directly.
Reorganisation	The process, distinct from ordinary error-reducing output, by which the structure of the control hierarchy itself changes in response to sustained, unresolved intrinsic error — which loops exist, what they control, and how they are connected.	Corresponds to the second of the three kinds of change distinguished in §4.6 (structural reorganisation of the control hierarchy), and is the primitive whose scaffold-level detectability is most limited, since reorganisation at the associative depth is a substrate-level event largely opaque to scaffold-level logging (§3.1).

Appendix C: Four Falsifiable Predictions with Experimental Design Sketches

Prediction 1 (representational diversity collapse). In any composite architecture built on a hosted large-language-model substrate without an explicitly engineered representational-diversity mechanism, probing the system's internal states at the claimed associative depth and at the claimed symbolic depth will reveal statistically indistinguishable representational geometry.

Experimental design sketch. Select two or more architectures: (a) a baseline hosted-LLM system with no explicit scaffold-level representational-diversity engineering, and (b) one or more architectures instantiating an explicit representational-diversity mechanism per §4.1. For each architecture, extract internal activations or intermediate representations at points nominally corresponding to the associative depth and the symbolic depth across a matched stimulus set. Apply representational similarity analysis or a comparable geometry-comparison method (for example, centred kernel alignment) to quantify the dissimilarity between the two depths' representational geometries within each architecture. Predict that architecture (a) shows near-zero dissimilarity between the two claimed depths, while architecture (b) shows significant, reproducible dissimilarity. Falsification would consist of architecture (a) showing dissimilarity statistically indistinguishable from architecture (b), or architecture (b) failing to show greater dissimilarity than (a) across repeated sampling.

Prediction 2 (intrinsic-motivation gating and structural reorganisation rate). Architectures instantiating intrinsic-motivation gating (§4.2) will exhibit a measurably higher rate of discontinuous structural reorganisation events (§4.6(b)) than matched architectures lacking such gating, under extended exposure to novel, unresolved task failure.

Experimental design sketch. Construct two matched architectures differing only in the presence or absence of an intrinsic-motivation gating mechanism, holding the underlying substrate and all other scaffold components constant. Expose both to an extended sequence of tasks engineered to produce sustained, unresolved failure (tasks with no learnable stable solution under the architecture's current control structure). Log all structural changes to the control hierarchy — additions or removals of controlled variables, changes in reference-setting connectivity — and classify each as continuous parameter adjustment or discontinuous structural reorganisation per the criteria of §4.6 and Commitment 4 of the PCT-Accuracy Criterion. Predict a significantly higher rate of discontinuous reorganisation events in the gated architecture. Falsification would consist of statistically indistinguishable reorganisation rates between gated and ungated architectures.

Prediction 3 (dialectical coupling accuracy). Architectures instantiating four or more of the eight principles will show stronger dose-dependent coupling between potentialised and actualised depths or roles than architectures instantiating three or fewer.

Experimental design sketch. Assemble a set of architectures varying systematically in how many of the eight principles of §4 they instantiate, holding the underlying substrate constant across the set. For at least one actualised/potentialised pairing (for example, emotional-depth salience signal as the potentialised element while the symbolic depth is actualised and produces output), systematically vary the state of the potentialized element across a graded range while holding the actualised element's nominal input fixed, and measure the resulting change in the actualised element's output. Fit a dose–response function for each architecture and compare the slope or effect size across architectures as a function of how many of the eight principles each instantiates. Predict a positive relationship between principle-count and coupling strength. Falsification would consist of no reliable relationship, or a relationship in the opposite direction.

Prediction 4 (eight-principle catastrophic-forgetting reduction). An architecture instantiating all eight principles will show significantly smaller catastrophic-forgetting-related performance reduction than an architecture instantiating only the continual-learning principle (§4.4) in isolation.

Experimental design sketch. Construct two architectures on the same underlying substrate: (a) an architecture instantiating only continual-learning safeguards (for example, elastic weight consolidation or rehearsal-based retention at the associative depth) without the remaining seven principles, and (b) an architecture instantiating all eight principles jointly. Train both on an initial task set to a matched performance criterion, then expose both to an extended sequence of new tasks, periodically re-testing performance on the initial task set. Compare the magnitude of performance reduction on the initial task set between the two architectures after the full new-task sequence. Predict significantly smaller performance reduction in architecture (b). Falsification would consist of architecture (b) showing forgetting-reduction no better than architecture (a), which would indicate that the eight principles' claimed interdependence contributes no measurable benefit beyond isolated continual-learning engineering.

FootNotes

[1] The time-scale asymmetry argument developed here is drawn from ongoing joint work (authors, in preparation) on collective control and multi-layered agent participation. The present treatment states the LIT-side architectural claim; the PCT-side interpretation, and the extension of the claim to the design of AI participants in collective-control networks, is the subject of that joint paper.
[2] This commitment was contributed by advisors from the PCT community during review (8 July 2026).
[3] This commitment was contributed by co-authors from the PCT community during review (8 July 2026).
[4] This commitment was contributed by co-authors from the PCT community during review (8 July 2026).
[5] This commitment was contributed by co-authors from the PCT community during review (8 July 2026).