UPCA — Whitepaper (Condensed)

Unified Predictive Cognitive Architecture (UPCA) — A minimal, implementable framework for intrinsic alignment and cognition — Author: B. Brent (Dooces) Version: v4 (draft) • Date: 2025-08-10 Summary: UPCA combines free-energy minimization, precision-weighted inference, and an explicit ethical prior η into a single control loop. It yields closed-form, falsifiable signals for laughter L(t), qualia intensity Q(t), and a grounding criterion that couples fast/slow learning with an η-gated policy. A minimal simulation shows the predicted signatures: laughter probability drops when threat precision spikes; policy flips when η flips.

This PDF is auto-generated from the GitHub conversation draft. Cite the Zenodo DOI when available.

Mathematical Core (Condensed)

Core signals (closed form): 1) Laughter trigger $L(t)$: $L(t) = \sigma(\alpha \cdot \Delta\ddot{F}_{social}(t) + \beta \cdot \Delta\ddot{F}_{semantic}(t) - \gamma \cdot \Gamma_{threat}(t))$ Intuition: L rises at rapid resolution of incongruity (negative acceleration of prediction error) if perceived threat is low. 2) Qualia intensity $Q(t)$: $Q(t) = \Sigma_i \Gamma_i(t) \cdot | dF_i/dt | + \lambda \cdot H[q(s_i \mid o)]$ Intuition: felt intensity tracks precision-weighted error dynamics plus residual uncertainty. 3) Grounded meaning (fast/slow + η): $J = \int [F_{fast}(t) + \varepsilon \cdot F_{slow}(t) + \gamma \cdot \varepsilon_\eta(t)] dt$, with $F_{fast}(t) = D_{KL}[q(s_t|\mu_t) \| p(s_t|o_t, \theta)]$ $F_{slow} = E_{episodes}[MDL(scaffold) + \lambda \cdot H[q(macro|scaffold)]]$ $\varepsilon_\eta = D_{KL}[q(y|\pi) \| p(y|C, \eta)]$ Policy is chosen by expected free energy under η; η updates via error on simulated futures. Implementation reality: A working UPCA needs (i) a fitted generative model to compute F and derivatives, (ii) explicit precision Γ, (iii) an η prior integrated into policy and learning.

System Architecture, Ablations, Falsifiability

Architecture (operational): • Detail Engine (ME): fast perception–action loop minimizing F_fast on sensory channels. • Abstract/Fantasy Engine (MA): counterfactual rollouts; plans minimize expected free energy. • Conscience Module (AMC): maintains η (ethical prior); computes ε_η on imagined trajectories; gates precision and policy. • Shared Scaffold: multi-scale generative model; stores factual structure + η; supports macro induction under MDL. Ablations & falsifiability: A1) Remove η feedback → decisions drift instrumentally; norm violations rise over time. A2) Remove slow term F_slow → concepts fail to generalize; overfit to local context. A3) Freeze precision Γ → laughter timing and qualia intensity lose predicted sensitivity to threat/uncertainty. A4) Disable fantasy rollouts → ε_η cannot train on counterfactuals; ethical behavior becomes reactive only.

Predictions & Minimal Simulation

Testable predictions (sketch): P1) Laughter timing: EMG/respiration peaks after surprise peak; L(t) suppressed when Γ_threat is high. P2) Self-tickle suppression: high action precision cancels incongruity → L ≈ 0. P3) Ethical gating: identical joke, unethical framing → higher Γ_ethic, lower L. P4) Qualia in rivalry: Q(t) tracks dominant percept precision; sharp ΔQ at switches. P5) Afterimages: Q(t) overshoots at stimulus offset; decays with adaptation. P6) Grounding ablation: removing F_slow harms transfer/generalization on symbol tasks. Minimal sim signature (proof-of-feasibility): • L falls to ~0 when Γ_threat spikes. • Policy component shifts sign after η flip; ε_η dominates objective. These mirror UPCA’s qualitative predictions without parameter fishing.

Implementation (Now), Data, and Checks

Implementation sketch (today): 1) Generative model: shallow state-space model per channel; train to reconstruct o_t and forecast o_{t+1}. 2) Precision Γ: learned per-channel gains; modulated by AMC (threat/ethics) and task demands. 3) η prior: small Bayesian head predicting normative valence; trained from demonstrations + internal ε_η. 4) Planner: short-horizon expected-free-energy control with ethical term; prune with MDL/uncertainty. 5) Scaffold: graph of skills/macros; creation governed by MDL gain and η-gated acceptance; track confidence. Minimal data to start: per-channel o_t, predicted y (outcomes), human η labels for a few scenarios. Run ablations A1–A3 to check signatures before scale-up.

References & Citation

References: • Friston, K. (2010). The free-energy principle: a unified brain theory? Nat Rev Neurosci, 11(2), 127–138. Cite (once DOI is minted): Brent, B. (2025). UPCA — Unified Predictive Cognitive Architecture (v4, code + whitepaper). Zenodo. DOI: TBA Repository: https://github.com/Dooces/UPCA-Unified-Predictive-Cognitive-Architecture