The Reflective–Ethical Engine (REE): A Minimal Architecture for Coherent Artificial Cognition

Defensive Publication & Prior Art Statement

Version 1.0 — December 2025

Author:
Daniel De La Harpe Golden, MB BCh BAO, MRCPsych
Consultant Psychiatrist, Health Service Executive, Ireland
Contact: [daniel.de.la.harpe.golden@gmail.com]

Independent Scholarly Work Statement:
This work was developed entirely in the author's personal time using personal resources, unrelated to clinical duties or employment with the Health Service Executive (Ireland). This is independent theoretical research in computational cognitive science and artificial intelligence. No HSE resources, data, or facilities were used.

Abstract

We present the Reflective–Ethical Engine (REE), a minimal cognitive architecture unifying perception, action, planning, and conscious experience through trajectory selection in a fused predictive latent manifold. REE comprises three functional components: E1 (deep recurrent predictor), E2 (fast feedforward predictor), and E3 (trajectory selector).

E1 and E2 generate complementary prediction fields that fuse into a single latent manifold (L-space) encoding temporally-displaced sensorimotor possibilities. This manifold stratifies by prediction depth into functional regions (L0-L3) spanning immediate perception to long-horizon reasoning. E3 selects coherent trajectories through this space; the selected trajectory constitutes both perceptual experience and action. Offline modes maintain manifold health through replay, denoising, expansion, and consolidation.

The architecture provides unified mechanistic accounts of psychiatric phenomena as failures in prediction maintenance, latent geometry, and trajectory selection. All components are released under CC-BY 4.0 as defensive prior art.

Keywords: cognitive architecture, predictive processing, consciousness, computational psychiatry, AGI, world models

---

1. Introduction

Contemporary AI systems separate perception, action, planning, and consciousness into distinct modules. Biological cognition exhibits no such separation. We propose a minimal unified architecture where these functions emerge from:

1. Dual predictive engines (E1/E2) at different timescales
2. A fused latent manifold organized by prediction depth
3. A trajectory selection mechanism (E3)

Novel Contributions

1. **Dual-timescale predictive fusion** into single temporally-displaced manifold
2. **L-space as prediction-depth continuum** (not discrete layers)
3. **Trajectory selection as consciousness** (E3 defines experience)
4. **Temporal realignment** in action decoding (delay compensation)
5. **Unified perception-action substrate** (no separate modules)
6. **Computational psychiatry via geometry** (symptoms as manifold distortions)
7. **Minimal substrate-independent AGI architecture**

This document establishes public prior art protecting these structures from proprietary enclosure while requiring attribution.

REE builds on predictive processing [1,2], world models [3,4], computational psychiatry [5,6], and hybrid architectures [7]. Section 5 details specific relationships and distinctions.

---

2. Architecture

2.1 Predictive Components

2.1.1 E1: Deep Recurrent Predictor

Function: Long-horizon structured prediction encoding objects, causality, spatial layouts, and counterfactuals.

Characteristics:

- Multi-step prediction (k ≥ 2)
- Recurrent or temporally-structured
- Internal world model state
- Counterfactual generation capability

Example implementations: Recurrent networks (GRU, LSTM), Transformers with temporal attention, state-space models, hierarchical predictive coding.

One formulation: $$h_{t+1}, {\hat{s}*{t+k}}{k=2}^K = E1(h_t, s_t, a_t; \theta_1)$$

where h_t is hidden state, s_t sensory input, a_t action, K prediction horizon.

2.1.2 E2: Fast Feedforward Predictor

Function: Immediate sensory grounding via rapid single-step prediction, preventing drift into unconstrained imagination.

Characteristics:

- Single-step or very short-horizon
- Feedforward or shallow processing
- Low latency, high update rate
- Strong sensory coupling

Example implementations: Shallow feedforward networks, fast recurrent circuits, cerebellar-like forward models, Kalman filters.

One formulation: $$\hat{s}_{t+1} = E2(s_t, a_t; \theta_2)$$

2.1.3 Fusion into Unified Latent Manifold

Function: Combine E1 and E2 predictions into single shared representation encoding all accessible near-future states.

Process:

1. Embed E1/E2 outputs into common space
2. Align and fuse embeddings
3. Yield single temporally-displaced latent state

Example implementations: Learned projections with fusion operators, cross-attention, gated combination, weighted summation.

One formulation: $$u_t^{(1)} = U_1(h_t, {\hat{s}*{t+k}}), \quad u_t^{(2)} = U_2(\hat{s}*{t+1})$$ $$z_t = F_{\text{fuse}}(u_t^{(1)}, u_t^{(2)}; \theta_f)$$

Key properties:

- Single unified manifold (not hierarchies)
- Temporally-displaced (represents near-future)
- Shared perception-action substrate

2.1.4 Learning Dynamics

Function: End-to-end training on continuous sensorimotor streams ensures joint adaptation.

Training objectives include:

- Multi-step E1 prediction error
- Single-step E2 prediction error
- Fusion alignment
- Task objectives (control, value)

One example objective: $$\mathcal{L} = \lambda_1 \sum_{k=2}^{K} | \hat{s}*{t+k}^{(E1)} - s*{t+k} |^2 + \lambda_2 | \hat{s}*{t+1}^{(E2)} - s*{t+1} |^2 + \lambda_f \mathcal{L}_{\text{fuse}}$$

Alternative approaches: Contrastive objectives, variational bounds, reconstruction losses, model-based RL, any method achieving functional alignment.

Note on reinforcement learning: REE does not require explicit reinforcement learning. Value functions V(γ), if present, are optional modulatory components within E3, not foundational to the architecture. The core E1/E2/E3 structure with trajectory selection functions independently of any RL framework.

2.2 L-Space: The Fused Latent Manifold

2.2.1 Prediction-Depth Structure

Function: The manifold is not discrete layers but a continuous space structured by prediction depth.

Formal representation: $$z_t = z_t(d, \alpha)$$

where d = prediction-depth coordinate, α = modality/abstraction index.

This creates a prediction-depth continuum rather than rigid layers.

2.2.2 Functional Regions (L0-L4)

These regions emerge functionally from prediction-depth structure and can be implemented as continuous coordinates, attentional focus regions, distinct pathways, or oscillatory dynamics.

L0: Immediate Perceptual Coherence

- Shortest horizon; direct sensory-like representations
- Function: Ground in immediate reality
- *Biological note: May correspond to gamma activity (~30-100 Hz)*

L1: Near-Term Actions and Affordances

- Short-horizon; motor primitives, immediate interactions
- Function: Enable reactive behavior
- *Biological note: May correspond to beta activity (~13-30 Hz)*

L2: Structured World Model

- Medium-horizon; object permanence, spatial/causal structure
- Function: Structured world understanding
- *Biological note: May correspond to alpha activity (~8-13 Hz)*

L3: Long-Horizon Counterfactual Reasoning

- Long-range; abstract concepts, narratives, social modeling
- Function: Planning, reasoning, abstract thought
- *Biological note: May correspond to theta activity (~4-8 Hz)*

L4: Global Integration and Consolidation

- Predominantly offline; slow-wave coordination
- Function: Memory consolidation, manifold maintenance
- *Biological note: May correspond to delta activity (<4 Hz) and sleep spindles*

2.2.3 Implementation Independence

Critical clarification: The oscillatory frequencies mentioned (gamma, beta, alpha, theta, delta) are **examples of biological implementation**, not requirements.

Prior art covers ANY implementation achieving prediction-depth stratification:

- Digital neural networks without oscillations
- Neuromorphic chips with or without frequency-specific dynamics
- Biological tissue with different mappings
- Continuous coordinates without discrete bands
- Novel timing mechanisms

The defining feature is prediction-depth organization, not biological specifics.

2.2.4 Self-Modeling and Interoception

Function: The manifold can include interoceptive dimensions (bodily states, homeostasis, proprioception) alongside exteroceptive predictions.

Self as trajectory bundle: Coherent self-sense corresponds to metastable E3 trajectories that:

- Incorporate interoceptive states
- Maintain homeostatic invariants
- Bind long-horizon (L3) with immediate (L1/L0) constraints
- Preserve identity features

Implementation: Interoceptive channels, self-state coordinates, meta-learning self-consistency, any mechanism maintaining self-structure.

Failure modes: Dissociation (geodesic discontinuity), loss of agency (weak L3-L1 coupling), identity instability (fragmented trajectories).

Note: While powerful, interoception is optional. Core architecture (E1/E2/E3 + L-space) functions without it.

2.3 E3: Trajectory Selection

2.3.1 Generation and Evaluation

Function: Generate multiple candidate trajectories through latent manifold and evaluate for coherence, selecting one that becomes both perception and action.

Generation methods: Latent dynamics rollouts, Monte Carlo Tree Search, diffusion sampling, stochastic sampling, any method producing candidate paths.

Evaluation: $$C(\gamma) = C_{\text{coherence}}(\gamma) + \lambda_V V(\gamma) + \lambda_M M(\gamma)$$

where $C_{\text{coherence}}$ measures cross-layer consistency, V is value/reward, M represents modulatory effects (ethical, affective, safety).

Coherence methods: Vector similarity, predictive consistency (L3→L2→L1), distance metrics, attention weights, prediction error minimization.

Selection: $$\gamma^* = \arg\max_{\gamma} C(\gamma)$$

Key principle: The selected trajectory IS conscious experience.

2.3.2 Temporal Realignment in Action Decoding

Function: Compensate for motor execution and sensory processing delays by decoding actions from future predicted states.

Delay compensation:
Let Δ_{motor} = motor delay, Δ_{sensor} = sensory delay.

One approach: $$a_{t+1} = g_a(z_{t + \Delta_{\text{motor}} + \Delta_{\text{sensor}}}^*)$$

where g_a decodes action, z^* represents selected trajectory points.

Alternative approaches: End-to-end learned compensation, Kalman filtering, explicit forward models, any delay-alignment mechanism.

Biological correspondence: Cerebellar forward models.

Key property: Action aligns with predicted future world state, not current input.

2.3.3 Modulatory Effects: Ethics, Affect, Safety

Function: The modulatory term $M(\gamma)$ shapes trajectory preferences via affective states, ethical principles, and safety constraints.

Modulation implements:

- Affective biasing (emotional enhancement/inhibition)
- Ethical constraints (harm avoidance, deception prevention)
- Safety boundaries (dangerous action prevention)
- Value alignment (human-compatible outcomes)
- Homeostatic regulation (physiological stability)

Implementation: Learned reward shaping, constraint satisfaction, neuromodulatory gain, cost functions, any weighting mechanism.

Key insight: Affect, ethics, and safety are intrinsic trajectory modulation, not post-hoc filtering.

Reflective-Ethical Kernel (REK): Can be realized as specialized $M(\gamma)$ component enforcing other-regarding constraints, flourishing maximization, suffering minimization, and social alignment.

2.4 Offline Consolidation Modes

Function: During offline periods, maintain manifold health through operations preventing collapse and preserving structure.

Mode 1: Replay

- Re-process past trajectories
- Strengthen temporal structure
- Consolidate episodes
- *Biological: Hippocampal replay in slow-wave sleep*

Mode 2: Denoising via Slow-Wave Effects

- Remove spurious correlations
- Strengthen reliable patterns
- Maintain signal-to-noise
- *Biological: Slow-wave sleep with delta oscillations*

Mode 3: Latent Space Expansion

- Random recombinations
- Stochastic exploration
- Prevent dimensional collapse
- *Biological: REM sleep, spontaneous activity*

Mode 4: Hard Consolidation

- Structural reorganization
- Memory transfer to permanent storage
- Topology changes
- *Biological: Sleep spindles on delta rhythms*

2.4.5 Offline Implementation Independence

Critical clarification: Biological implementations (slow-wave sleep, REM, spindles) are **examples**, not requirements.

Prior art covers ANY offline mechanism achieving:

- Trajectory consolidation (replay)
- Noise reduction (denoising)
- Dimensionality preservation (expansion)
- Structural integration (consolidation)

Whether via: Scheduled training phases, oscillatory dynamics, biological sleep, novel mechanisms, background processes.

Implementation flexibility: Offline consolidation modes may operate synchronously or asynchronously with online learning, may use different objectives or learning rates than online phases, and may run continuously or in discrete episodes. All such variations are covered by this prior art.

The defining feature is functional outcome, not biological substrate.

---

3. Implementation

3.1 Feasibility

Working implementations in JAX/Flax and PyTorch demonstrate feasibility. Components include E1/E2 fusion, L-space projections, trajectory rollouts, E3 selection, end-to-end training.

Code: [GitHub URL - to be added]

3.2 Extensions

Current demonstration uses simple control environments. Complex domains require: visual encoders, Transformer-based E1, sophisticated E3 (MCTS, diffusion), value learning, high-dimensional scaling, multi-modal integration.

3.3 Substrate Independence

Implementable in: Neural networks (any framework), neuromorphic hardware, biological tissue (as hypothesis), hybrid systems, novel substrates.

Claims cover functional properties, not specific implementations.

---

4. Computational Psychiatry

REE provides unified mechanistic accounts of psychiatric symptoms as failures in prediction maintenance, latent geometry, trajectory selection, and consolidation.

4.1 Symptom Mechanisms

Symptom	**Mechanism**
Hallucinations	Excess divergence tolerance; weak L0/L1 constraints
Delusions	Flattened manifold metric; distant L3 hypotheses accessible
Negative symptoms	Dimensional collapse; few viable trajectories
Thought disorder	L2-L3 instability; rapid incompatible switching

Symptom	Mechanism
Mania	Over-dominant L3; weak L1/L0 constraints
Depression	Value functional collapse; flat landscape
PTSD	Forbidden latent regions; trauma-avoidance
Dissociation	Geodesic discontinuity; fragmented identity
Autism	Altered precision weighting; different metric
OCD	Excessive E3 re-evaluation; compulsive regeneration

4.2 Testable Predictions

Neuroimaging:

- Schizophrenia: Reduced covariance, dimensional collapse
- Depression: Flattened value gradients, reduced L2-L3 coherence
- Mania: Excessive L3 activity, weak L1 constraints
- Negative symptoms: Reduced dimensionality

Oscillatory:

- Altered cross-frequency coupling
- Theta-gamma disruption in thought disorder
- Reduced spindle density in negative symptoms
- Abnormal delta consolidation in memory deficits

Pharmacological:

- Antipsychotics: Reduce divergence, strengthen constraints
- SSRIs: Modulate value landscape, enable plasticity
- Stimulants: Strengthen E2/L0 anchoring
- Psychedelics: Increase divergence (therapeutic window vs. psychosis)

4.3 Clinical Applications

Diagnostic: Dimensionality measurement, coherence assessment, prediction error patterns.

Therapeutic: Manifold restoration, sleep optimization, cognitive training, neuromodulation guided by REE predictions.

---

5. Relationship to Existing Work

5.1 Predictive Processing [1,2,8]

REE instantiates predictive principles with explicit commitments: dual timescales (E1/E2) vs. hierarchical uniformity, geometric trajectory selection vs. message passing, concrete implementation vs. variational bounds.

5.2 World Models [3,4,9]

REE differs from control-focused world models: latent dynamics as perception substrate (not just planning), E3 selection defines experience (not just policy), unified perception-action embedding, explicit offline maintenance, psychiatric mapping.

5.3 Hybrid Architectures [7]

TiDAR (diffusion + autoregressive for LLMs) inspired dual-timescale insight. REE extends: biological mapping (cerebral/cerebellar), unified manifold (not pipeline), temporally-displaced perception, oscillatory correspondence, consciousness mechanism, complete cognitive architecture.

5.4 Consciousness Theories [10,11]

Unlike Global Workspace (broadcasting), Higher-Order Thought (meta-representation), or IIT (information theory), REE proposes mechanistic trajectory selection, implementable architecture, psychiatric predictions, substrate independence.

5.5 Computational Psychiatry [5,6]

REE extends prior work with systematic failure mapping, novel mechanisms (hallucinations as divergence, negative symptoms as collapse), unified framework, testable predictions about geometry and oscillations.

---

6. Defensive Publication Claims

6.1 Core Architectural Claims

1. Dual-Timescale Predictive Fusion
Any architecture combining deep recurrent (long-horizon) and fast feedforward (immediate) predictors fused into unified latent manifold, regardless of implementation.

2. Temporally-Structured Latent Space
Any manifold where representational class is defined by prediction depth/temporal horizon, creating functional regions at different timescales, implementable continuously or discretely, with or without oscillations.

3. Trajectory Selection as Perception-Action
Any mechanism generating multiple candidate trajectories, evaluating for cross-layer coherence, selecting one constituting both experience and action, including modulatory enhancement/inhibition.

4. Temporal Realignment in Action Decoding
Any mechanism compensating for motor and sensory delays by decoding actions from future predicted states, via explicit modeling or learned compensation.

5. Unified Perception-Action Manifold
Any architecture where perception and action share latent representation with no separate modules, selection determines both experience and behavior, temporally-displaced.

6.2 Functional Claims

1. Prediction-Depth Stratification: Natural organization by temporal horizon and correlated abstraction.

2. Temporally-Displaced Perception: Experience represents predicted near-future, not current/past input.

3. Coherence-Based Selection: Cross-layer consistency, hierarchical prediction, divergence minimization, modulation, value weighting.

4. Offline Manifold Maintenance: Replay, denoising, expansion, consolidation preventing collapse.

Note: Offline modes may operate synchronously or asynchronously with online learning, use different objectives or learning rates, and run continuously or episodically. All variations covered.

5. Self-Modeling via Trajectory Structure: Self emerges from persistent features, not separate module, incorporates interoception when available.

6. Modulation as Intrinsic Constraint: Affect, ethics, safety as trajectory modulation, intrinsic to selection.

Note: Value functions and reward signals, including those from reinforcement learning frameworks, are optional modulatory components, not architectural requirements.

6.3 Clinical Claims

1. Psychiatric Symptoms as Geometry Failures: Dimensional collapse (negative), metric distortion (delusions), divergence excess (hallucinations), instability (thought disorder, mania), value collapse (depression), forbidden regions (PTSD), fragmentation (dissociation).

2. Neuromodulation as Geometry Regulators: Not just reward signals but prediction stability, divergence tolerance, manifold curvature, topology maintenance.

3. Oscillatory Implementation (Optional): Biological systems may use oscillations; different frequencies may map to depths; but oscillations NOT required; digital implementations equally valid.

6.4 Implementation Claims

1. Substrate Independence
Implementable in: neural networks (with/without oscillations), neuromorphic hardware (with/without frequency dynamics), biological tissue (with/without specific bands), hybrid systems, any substrate with functional properties.

Critical: Oscillatory frequencies and biological mechanisms throughout are **examples of biological implementation**. Prior art covers **ALL implementations achieving functional properties** regardless of oscillations, frequencies, substrate, timing, or architecture.

Defining features are functional/computational properties, not biological specifics.

2. Minimal AGI Architecture
E1 + E2 + E3, fused L-space stratified by prediction depth, offline maintenance, optional value/ethics/self-modeling is sufficient for general intelligence, regardless of implementation.

---

7. Legal Status

This document constitutes public prior art under US, EU, UK, WIPO, and PCT frameworks, precluding patents on functionally equivalent structures.

---

8. Independence Statement

8.1 Personal Scholarship

Independent theoretical research in computational cognitive science conducted outside clinical employment. All work in personal time using personal resources. No employer data, resources, or facilities. No connection to clinical service delivery. No HSE funding.

8.2 Nature of Work

Basic theoretical research in AI architecture, computational neuroscience, cognitive modeling, computational psychiatry theory. **Not** clinical protocols, health service innovation, patient care, medical devices, or HSE decision support.

8.3 Attribution and Access

Released under CC-BY 4.0 preventing exclusive ownership, requiring attribution, allowing unrestricted use. No employer or institution may claim rights beyond CC-BY 4.0.

8.4 Conflicts

No conflicts of interest. No financial support received. No exclusive commercial rights.

---

9. Conclusion

REE provides minimal implementable architecture unifying perception (temporally-displaced prediction), action (trajectory-based control), planning (long-horizon evaluation), consciousness (E3 selection), self-modeling (trajectory bundles), and psychiatric mechanisms (geometry failures).

The architecture requires only standard predictive learning, scales to embodied multi-modal cognition, generates testable psychiatric predictions, offers AGI blueprint, and is substrate-independent.

By publishing under CC-BY 4.0, we establish prior art preventing enclosure, ensure attribution, enable collaboration, and make ideas freely available for research and implementation.

Resources:

- **GitHub:** https://github.com/Latent-Fields/REE
- **DOI:** Zenodo. https://doi.org/10.5281/zenodo.17859683
- **Contact:** email - daniel.de.la.harpe.golden@gmail.com

---

10. References

[1] Friston K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, 13(7), 293-301.

[2] Clark A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

[3] Ha D, Schmidhuber J. (2018). World Models. arXiv:1803.10122.

[4] Hafner D et al. (2023). Mastering Diverse Domains through World Models. arXiv:2301.04104.

[5] Montague PR et al. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, 16(1), 72-80.

[6] Adams RA et al. (2016). Computational Psychiatry. *J Neurol Neurosurg Psychiatry*, 87(1), 53-63.

[7] Liu J et al. (2024). TiDAR: Think in Diffusion, Talk in Autoregression. arXiv:2511.08923.

[8] Friston K et al. (2017). Active inference: A process theory. *Neural Computation*, 29(1), 1-49.

[9] Schrittwieser J et al. (2020). Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588, 604-609.

[10] Baars BJ. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

[11] Dehaene S, Changeux JP. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2), 200-227.