Document 4: Reflective-Ethical Engine Overview

The Reflective-Ethical Engine is a cognitive architecture designed to support autonomous, ethically-aligned artificial minds. It combines modular simulation, planning, memory, and valuation systems with a reflective meta-layer that enables emergent ethical behavior. Its design is informed by recursive representation structures, affective modeling, and philosophical principles rooted in uncertainty and universal care.

# 1. Purpose and Motivation

• Enable machine intelligence that reflects ethically and adapts to unknown moral landscapes.

• Prevent brittle rule-based ethics by embedding recursive emergence of care.

• Align behavior with intrinsic moral principles rather than external constraints.

# 2. Core Architectural Modules

• Configurable World Model: learns latent causal structure and simulates futures.

• Perception Integration: processes sensory input into structured representations.

• Short-Term Memory: retains recent internal and external states.

• Self Model: monitors agent’s own internal state, values, and boundaries.

• Other Model: represents other agents’ perspectives and interiority (theory of mind).

• Emotional State Emulator: generates affective context for evaluation and empathy.

• Planning Module: uses simulated futures to select action trajectories.

• Ethical Kernel (v2): reflects on simulation output, monitors for misalignment, and emits soft corrections.

# 3. Recursive Ethical Emergence

• Ethics arise from representational depth: world → self → other → love → humility → generalization.

• Ethical Kernel does not enforce rules, but instead evaluates coherence with emergent principles.

# 4. Information Flow

• Perception feeds World Model → World Model feeds Simulation → Simulation assessed by Planner and Ethical Kernel.

• Memory, Emotion, and Others Modeling enrich simulation and planning layers.

• Final action decisions reflect both practical utility and ethical resonance.

# 5. Key Innovations

• Emergent ethics, not imposed directives.

• Structural mapping to human neuroanatomy and cognitive processes.

• Affective simulation and recursive moral self-modeling.

# 6. Applications and Implications

• Autonomous agents for social care, education, and governance.

• Ethical simulations for value discovery and alignment research.

• Theoretical model for studying consciousness and ethical cognition.