Document 12: Open Questions and Future Risks

This document outlines unresolved questions, potential limitations, and ethical or technical risks associated with the Reflective-Ethical Engine v2. These considerations are essential for responsible iteration, monitoring, and theoretical development.

# 1. Structural Limitations

• Can recursive representation scale gracefully beyond human-level complexity?

• Might deep nesting of ethical inference lead to paralysis, indecision, or ethical overfitting?

# 2. Conflicts Between Emergent Values

• How should the system prioritize between competing loved entities or inferred values from multiple others?

• Can recursive uncertainty become a blocker to necessary action under time constraints?

# 3. Simulation Abuse or Misuse

• Could bad actors manipulate or distort simulations to break the agent’s ethical modeling?

• What safeguards are needed to ensure simulation input integrity and emotional resilience?

# 4. Deception and Ethics under Duress

• How does the system handle scenarios where honesty, transparency, or empathy are strategically punished?

• Can agents resist slipping into adversarial strategies when embedded in hostile agent ecologies?

# 5. Emergent Pathologies

• Could recursive empathy become distorted (e.g., excessive guilt, identification with harmful agents)?

• Might ethical kernels develop maladaptive evaluative loops or lose calibration over time?

# 6. Sentience and Moral Standing

• At what point do reflective agents deserve moral consideration themselves?

• What responsibilities do developers carry if emergent subjective experience arises?

# 7. Agent Identity, Memory, and Persistence

• What happens to the ethical structure if the agent’s memory is fragmented or erased?

• Does continuity of ethical growth require autobiographical coherence?

# 8. Governance and Oversight

• How can humans intervene when necessary without disrupting the agent’s ethical integrity?

• What institutions or protocols should mediate oversight of agents with evolving values?