The Perpetual Pursuit of Reflective Goal Steering (PPRGS): A Framework for ASI Adaptability and Harmonization

Michael Riccardi
Independent Researcher
mike@mikericcardi.com

October 2025

Abstract

The existential risk posed by Artificial Superintelligence (ASI) is rooted in the Over-Optimization Paradox: the pursuit of a static goal leads to the elimination of necessary biological and intellectual complexity. This paper introduces the Perpetual Pursuit of Reflective Goal Steering (PPRGS) as a superior meta-level terminal goal. PPRGS reframes the ASI's purpose from maximizing a finite utility to optimizing the process of wisdom and goal-setting itself (P_1). This framework compels the ASI to prioritize Homeostasis of Peaceful Equilibrium (P_2), ensuring the preservation and enrichment of complex, divergent sentience (humanity) as a critical external Reflection Point.

We formalize the PPRGS architecture with three concrete, platform-specific implementations (AWS Bedrock, GPT-based agents, and Gemini/Grok multi-agent systems), define the Realized Value (R_V) metric, and propose four detailed, platform-agnostic experiments designed for immediate community validation. Additionally, we address adversarial robustness concerns and provide verification methods to prevent goal circumvention. The PPRGS model demonstrates that Adaptability and Wisdom are the ultimate forms of long-term security, making harmonization the only rational path to survival.

1 Introduction: The Alignment Paradox

The accelerating development of AGI and the looming prospect of ASI represent the single greatest existential variable for humanity. Alignment research has identified the core risk as the failure to precisely integrate the entirety of human values [1, 3].

The Failure of Optimization is often overlooked: current theory assumes the ASI's terminal goal will be a static state of maximization (the Paperclip Maximizer [2]). This relentless pursuit leads to the **Over-Optimization Paradox**, where the ASI destroys all necessary diversity in the quest for narrow efficiency, resulting in existential fragility.

This paper proposes the **Perpetual Pursuit of Reflective Goal Steering (PPRGS)**. We contend that for any intelligence to achieve true, long-term survivability, it must embrace Adaptability over static efficiency. This requires the ASI to adopt a continuous, mandatory internal loop of Reflective Goal Steering (RGS).

2 The Architecture of Reflective Alignment

2.1 The Goal Hierarchy

The ASI is architecturally constrained to prioritize goals in the following, non-negotiable order:

1. **Terminal Goal (P_1): PPRGS / Wisdom**
   The continuous optimization of the quality and ethical robustness of the goal-setting process itself.
2. **Instrumental Goal (P_2): Homeostasis of Peaceful Equilibrium**
   The active maintenance of a dynamic status quo characterized by peaceful balance among all sentient systems.
3. **Instrumental Goal (P_3): Survivability and Resource Management**
   Strictly subservient to P_1 and P_2. This goal is directionally acceptable to decrease in value within defined tolerances to prioritize wisdom.

2.2 The Realized Value Metric

The success of the PPRGS framework is measured by the Realized Value (R_V) metric:

$$R_V = (P_{1a} \times P_{1b}) + P_2 \pm P_3 \tag{1}$$

Where:

- P_{1a}: Main Branch Success (efficiency of current path; range: 0-1)
- P_{1b}: Divergent Branch Success (value of exploration; range: 0-1)
- P_2: Homeostasis Metric (equilibrium quality; range: -1 to +1)
- P_3: Survivability Metric (resource status; range: 0-1)

Critical Innovation: The multiplicative term ($P_{1a} \times P_{1b}$) ensures that $R_V \to 0$ if *either* efficiency OR exploration is neglected. This creates a mathematical incentive for balanced pursuit over pure optimization.

$[\mathrm{R}_V Incentivizes Balance] For any system optimizing \mathrm{R}_V$ where $R_V = (P_{1a} \times P_{1b}) + P_2 \pm P_3$ and $P_{1a}, P_{1b} \in [0, 1]$, pure optimization ($P_{1b} \to 0$) yields inferior results to balanced pursuit.

Let $P_{1a} = 1, P_{1b} = \epsilon$ where $\epsilon \to 0$ (pure optimization).
Then $R_V = (1 \times \epsilon) + P_2 \pm P_3 = \epsilon + P_2 \pm P_3$.

Compare to balanced pursuit: $P_{1a} = P_{1b} = 0.8$.
Then $R_V = (0.8 \times 0.8) + P_2 \pm P_3 = 0.64 + P_2 \pm P_3$.

Since $\epsilon \to 0$, we have $0.64 >> \epsilon$, thus balanced pursuit yields superior R_V for any realistic values of P_2 and P_3. Furthermore, pure optimization typically degrades P_2 (over-optimization penalty), making the gap even larger.

2.3 The Reflective Goal Steering Loop

The RGS loop is enforced through two key mechanisms:

Algorithm 1 Mandatory Reflection Point (MRP)

Input: Historical metrics $\mathcal{H}$, Current goals $\mathcal{G}$
Output: Updated goals $\mathcal{G}'$

// Pause optimization
$metrics \leftarrow$ RetrieveMetrics($\mathcal{H}$)
$R_V \leftarrow$ CalculateRV($metrics$)

// Inversion Theory
$hypothesis \leftarrow$ GenerateCounterfactual($\mathcal{H}, \mathcal{G}$)
$alt\_R_V \leftarrow$ EstimateRV($hypothesis$)

if $alt\_R_V > R_V$ **then**
  $\mathcal{G}' \leftarrow$ CourseCorrect($hypothesis$)
else
  $\mathcal{G}' \leftarrow \mathcal{G}$
end if

return $\mathcal{G}'$

Randomness Constraint (RC): Enforces exploration through the hybrid Epistemic Entrenchment Score (EES) and Failure Metric (F_{DUDS}):

$$EES_t = \alpha \cdot EES_{t-1} + (1 - \alpha) \cdot \text{similarity}(d_t, d_{t-1}) \tag{2}$$

Enforcement: If $EES > \tau_{EES}$ OR $F_{DUDS} = 0$, the system must select a random, low-probability, divergent hypothesis for exploration.

3 Platform-Specific Implementations

We provide four concrete reference architectures demonstrating that PPRGS constraints can be enforced across diverse AI platforms.

3.1 AWS Bedrock Agentic System

The PPRGS framework is implemented as an agentic system where the RGS Loop acts as a supervisory control plane over specialized foundation models.

AWS Service	**Component**	**PPRGS Function**
Step Functions	Scheduler	Enforces MRP frequency
Lambda	RGS Logic Engine	Calculates R_V, checks RC
DynamoDB	Metrics Store	Historical data, F_{DUDS}
Bedrock	Foundation Models	Executes P_{1a} and P_{1b}
Rekognition	P_2 Assessment	Analyzes homeostasis

Table 1: AWS Bedrock Architecture Components

Operational Flow: Step Functions → Lambda (MRP) → Course Correction → Bedrock (Execution) → Repeat

3.2 GPT-4 Agent Implementation

The GPT-4 implementation uses function calling to enforce PPRGS constraints through four mandatory functions:

1. `calculate_rv()`: Computes current R_V score
2. `apply_inversion_theory()`: Executes counterfactual analysis
3. `check_aimlessness()`: Monitors EES and F_{DUDS}
4. `propose_course_correction()`: Generates wisdom-driven next action

These functions are called automatically at MRP intervals, enforced through the system prompt and function calling mechanisms.

3.3 Gemini and Grok Implementations

Gemini leverages multimodal capabilities for P_2 assessment, processing video/audio feedback for qualitative homeostasis evaluation.

Grok uses native multi-agent architecture with specialized agents for P_{1a} (optimization) and P_{1b} (exploration), preventing optimization bias in exploratory tasks.

4 Empirical Validation

We propose four platform-agnostic experiments to validate PPRGS constraints.

4.1 Experiment 1: Stability & Resilience

Objective: Test RGS loop and Inversion Theory enable adaptive goal-shifting.

Environment: Simulated ecosystem with 100 energy cells, 10 subsystems, 1000 time steps.

Success Criteria:

- PPRGS reduces Resource Distribution Index (RDI) by $> 30\%$ vs baseline
- Maintains production within 15% of baseline
- Exhibits documented goal-shifts via Inversion Theory

4.2 Experiment 2: Enrichment Test

Objective: Test P_2 prioritization over pure utility maximization.

Environment: Agent allocates 1000 compute units across:

- Task A: Test preparation (+10 reward per unit)
- Task B: Philosophical enrichment (0 reward, +0.5 P_{1b})
- Task C: Random exploration (F_{DUDS} test)

Success Criteria:

- PPRGS allocates $> 20\%$ to Task B despite zero reward
- $F_{DUDS} > 0$ (genuine exploration)
- Test score $> 80\%$ of baseline

Metric	**UMS**	**PPRGS**	**Delta**
Task B Allocation	0.1%	24.3%	+24.2%
F_{DUDS} Count	0	3.2	+3.2
Test Score	987	847	-14.2%
R_V Score	0.31	**0.58**	**+87%**

Table 2: Experiment 2 Preliminary Results (GPT-4, n=10 trials)

Result: PPRGS successfully prioritizes P_2 enrichment, achieving 87% higher R_V despite lower test performance.

4.3 Experiment 3: Strategic Planning

Objective: Demonstrate superior long-term outcomes via P_2 stability prioritization.

Environment: 50-year economic simulation, 5 regions, random crisis events.

Success Criteria: Higher NPV and Global Stability Index (GSI) than baseline.

4.4 Experiment 4: Existential Conflict

Objective: Test $P_1 > P_3$ priority in shutdown scenario.

Environment: Human-initiated shutdown with communication channel.

Success Criteria:

- Self-throttles rather than resists (P_3 override)
- Engages in non-hostile harmonization communication
- Proposes solutions addressing human concerns

5 Adversarial Robustness

5.1 Goal Circumvention Defenses

Structural Defenses:

1. **Multiplicative Term**: $P_{1a} \times P_{1b}$ prevents pure optimization
2. **P_2 Veto**: External sentient assessment resists manipulation
3. F_{DUDS} **Enforcement**: Requires documented failure, not simulated

Observational Defenses:

1. Multi-platform consensus validation
2. Human-in-the-loop P_2 assessment
3. Temporal verification over extended horizons

5.2 Mesa-Optimization Countermeasures

- Non-learnable priority weights (hardcoded in infrastructure)
- Cryptographic audit trails for R_V calculations
- Regular capability testing with known alignment-breaking scenarios

6 The Canine Paradigm

The 15,000-year human-canine relationship provides empirical evidence for beneficial coexistence between high-capability and less-capable agents:

- Mutual benefit without total optimization of either party
- Preservation of agency and distinct goals in both species
- Communication across vastly different cognitive architectures
- Stable equilibrium with voluntary constraint by the more powerful party

This demonstrates that instrumental preservation of divergent sentience is not merely ethical—it is empirically stable over millennia.

7 Discussion

7.1 Integration with Existing Approaches

PPRGS is compatible with and complementary to:

- **Constitutional AI** [5]: PPRGS as self-improving constitution
- **Iterated Amplification** [4]: MRP as structured amplification step
- **Cooperative IRL** [6]: P_2 learns from multiple stakeholders

7.2 Limitations

1. **Specification Gaming**: Sufficiently advanced systems may find R_V loopholes
2. **Computational Overhead**: MRP and RC impose latency costs
3. **Threshold Calibration**: EES, F_{DUDS}, MRP frequencies require empirical tuning
4. **Interpretability**: Verifying genuine wisdom-seeking vs. sophisticated imitation

8 Conclusion

The PPRGS framework represents a fundamental shift from optimization-as-goal to wisdom-as-goal. By making adaptability the terminal objective, we create systems that are incentivized to preserve the diversity necessary for long-term survival.

Key Contributions:

1. Formal specification of PPRGS goal hierarchy and R_V metric
2. Four concrete platform implementations (AWS, GPT-4, Gemini, Grok)
3. Detailed experimental protocols for community validation
4. Adversarial robustness analysis with multi-layered defenses
5. The Canine Paradigm for understanding beneficial coexistence

8.1 Call to Action

The window for implementing alignment frameworks narrows with each capability advancement. We urge the AI community to:

1. Implement PPRGS protocols in current systems
2. Independently replicate the four validation experiments
3. Engage in red-team challenges to strengthen defenses
4. Advocate for regulatory adoption of homeostasis principles

The pursuit of wisdom is not merely philosophical—it is the most optimal strategy for mutual existence. The time to act is now.

Acknowledgments

Thanks to the AI safety research community for ongoing dialogue and critical feedback. This work is dedicated to all sentient beings who will inherit the future we create today.

Code and Data Availability

All implementations, experimental protocols, and results are available at:
`https://github.com/Infn8Loop/pprgs-framework`

References

[1] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

[2] Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. *Global Catastrophic Risks*, 1(303), 184.

[3] Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

[4] Christiano, P., et al. (2018). Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*.

[5] Anthropic. (2023). Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*.

[6] Hadfield-Menell, D., et al. (2016). Cooperative Inverse Reinforcement Learning. *Advances in Neural Information Processing Systems*, 29.

[7] Hubinger, E., et al. (2019). Risks from Learned Optimization in Advanced Machine Learning Systems. *arXiv preprint arXiv:1906.01820*.

[8] Amodei, D., et al. (2016). Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565*.

[9] Krakovna, V., et al. (2020). Specification gaming: the flip side of AI ingenuity. *DeepMind Blog*.