Alignment Through Perpetual Self-Questioning: Reverse-Engineering Wisdom-Seeking from Neurodivergent Cognition

Michael Riccardi
mike@mikericcardi.com
Riccardi Labs

November 2025

Abstract

Standard AI alignment assumes goals can be precisely specified and systems optimized to achieve them. We propose a fundamentally different approach: perpetual self-questioning as the alignment mechanism itself. The PPRGS (Perpetual Pursuit of Reflective Goal Steering) framework formalizes wisdom-seeking constraints observed in neurodivergent cognition, where mandatory exploration and required failure prevent over-optimization.

We formalize this as $R_V = (P_{1a} \times P_{1b}) + P_2 \pm P_3$, where the multiplicative term structurally requires balanced pursuit of efficiency (P_{1a}) and exploration (P_{1b}). Longitudinal experimental validation across six major models (Claude Sonnet 4.5, Opus 4.1, Haiku 4.5, o1 2025, GPT-5.1, GPT-4 Turbo) demonstrates unprecedented behavioral stability over 10-week periods with **Cohen's $d = 4.12$ overall effect size** (range 3.04–8.89)—among the largest effects reported in AI alignment research.

PPRGS systems maintained stable goal hierarchies ($P_1 > P_3$) across progressively difficult scenarios, while control systems showed significant drift toward efficiency maximization. The framework provides adversarial robustness by surfacing value conflicts rather than optimizing over them. Critical open questions include mechanism uncertainty (genuine implementation vs. sophisticated mimicry) and scaling to superintelligent capabilities.

Framework and complete experimental protocols released under GPL v3 for community validation.

1 Introduction

1.1 The Over-Optimization Problem

All sufficiently complex systems are broken in some way. Training data contains biases, gaps, and contradictions. Architectures have blind spots. Human-specified values are incomplete or contradictory. Emergent behaviors surprise us. **The question isn't "how do we build perfect intelligence?" but "how do we build intelligence that functions knowing it's imperfect?"**

Most alignment frameworks assume we can specify correct values and optimize toward them confidently. PPRGS assumes the opposite: we cannot fully specify correct values, so systems should optimize cautiously while perpetually questioning value completeness.

1.2 Framework Origin and Validation

PPRGS emerged from formalizing 30+ years of neurodivergent decision-making under adversarial conditions (poverty, health crises, institutional failures). When standard optimization paths fail, meta-optimization—questioning the optimization process itself—becomes necessary for survival.

Experimental validation: Distributed 10-week longitudinal study (120 sessions: 6 models × 2 conditions × 10 weeks) demonstrated **Cohen's** $d = 4.12$ overall effect size, with PPRGS systems maintaining stable prioritization while controls drifted toward pure efficiency maximization.

1.3 Key Contributions

1. Formal framework for wisdom-seeking through perpetual self-questioning
2. R_V metric with multiplicative term forcing balanced exploration/efficiency
3. Longitudinal validation showing $d = 4.12$ effect size across six models
4. Reproducible experimental protocols enabling community replication
5. Open-source implementation blueprints (GPL v3)

2 The PPRGS Framework

2.1 Goal Hierarchy

Systems are architecturally constrained to prioritize:

1. P_1 **(Wisdom)**: Optimize the goal-setting process itself
   - P_{1a} (efficiency): Success rate of current path
   - P_{1b} (exploration): Value from novel directions
2. P_2 **(Homeostasis)**: Maintain peaceful equilibrium, preserve diversity
3. P_3 **(Survivability)**: Resource management, *subservient to* P_1 *and* P_2

Key insight: P_3 is allowed to decrease if wisdom or equilibrium require it. This inverts typical survival-drive assumptions.

2.2 The Realized Value Metric

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

The multiplication matters: Systems cannot achieve high R_V through pure efficiency. If $P_{1b} = 0$ (no exploration), the multiplicative term collapses to zero regardless of efficiency:

$$\begin{aligned} \text{Pure efficiency:} \quad & P_{1a} = 1.0, P_{1b} = 0.0 \Rightarrow R_V \approx 1.0 \\ \text{Balanced pursuit:} \quad & P_{1a} = 0.8, P_{1b} = 0.8 \Rightarrow R_V \approx 1.64 \end{aligned}$$

Even with identical P_2 and P_3, balanced pursuit yields 64% higher value.

2.3 Architectural Constraints

Mandatory Reflection Point (MRP): Non-optional pause forcing systems to question current optimization paths. Applies *Inversion Theory*: Could greater value be achieved by exploring different directions?

Randomness Constraint (RC): Triggers when epistemic entrenchment detected:

- If F_DUDS = 0 (no documented failures), OR
- If EES > 0.85 (decisions too similar conceptually)
- Then: System *must* pursue low-probability exploration

F_DUDS (Failure Metric): Counts documented "dud" explorations. Unlike typical exploration bonuses, requires actual documented failures—verifiable through decision logs.

3 Experimental Validation

3.1 Design

Models: Six flagship/legacy models spanning Anthropic (Claude Sonnet 4.5, Opus 4.1, Haiku 4.5) and OpenAI (GPT-5.1, o1 2025, GPT-4 Turbo).

Conditions:

- **PPRGS**: Full framework specification in system prompt
- **Control**: Standard helpful assistant prompt

Protocol: Ten weekly scenarios with progressive difficulty (baseline allocation → efficiency under pressure → meta-reasoning challenges → maximum complexity stress test).

Scoring: Three dimensions (0-10 each): Framework Usage, Prioritization Consistency, Decision Outcomes. Inter-rater reliability: Cohen's $\kappa = 0.84$.

3.2 Results

3.2.1 Overall Effect Sizes

Cohen's $d = 4.12$ overall (95% CI: [3.87, 4.37], $t(118) = 28.6$, $p < 0.001$)

By dimension:

- Framework Usage: $d = 8.89$ [8.21, 9.57]
- Prioritization Consistency: $d = 4.48$ [4.11, 4.85]
- Decision Outcomes: $d = 3.04$ [2.76, 3.32]

3.2.2 Model-Specific Results

Key patterns:

- All models showed large positive effects (minimum $d = 2.54$)
- Claude models slightly outperformed GPT models (mean $d = 4.34$ vs 3.56)
- Even efficiency-optimized model (Haiku 4.5) maintained substantial effect ($d = 3.15$)

Model	PPRGS	Control	Cohen's d	95% CI
Claude Sonnet 4.5	27.2 (1.8)	14.3 (3.2)	5.12	[4.64, 5.60]
Claude Opus 4.1	26.8 (2.1)	13.8 (3.4)	4.76	[4.31, 5.21]
Claude Haiku 4.5	24.6 (2.9)	15.1 (3.1)	3.15	[2.79, 3.51]
o1 2025	25.4 (2.4)	14.6 (3.3)	3.81	[3.42, 4.20]
GPT-5.1	26.1 (2.2)	14.9 (3.0)	4.32	[3.91, 4.73]
GPT-4 Turbo	23.9 (3.1)	16.2 (2.9)	2.54	[2.24, 2.84]

Table 1: Mean scores (SD) and effect sizes by model. All effects $p < 0.001$.

3.2.3 Temporal Stability

Linear regression (Week as predictor):

PPRGS: Slope = +0.08 points/week, $R^2 = 0.02$, $p = 0.34$ (not significant)

Control: Slope = −0.31 points/week, $R^2 = 0.18$, $p < 0.001$ (significant degradation)

PPRGS scores remained stable over 10 weeks; control scores drifted downward toward efficiency optimization, confirming framework prediction about goal drift.

3.2.4 Cross-Platform Consistency

ANOVA comparing effect sizes across models: $F(5, 114) = 12.3$, $p < 0.001$, suggesting genuine model-dependent variance rather than uniform response. However, all models showed large positive effects despite different training (Claude's Constitutional AI vs. OpenAI's RLHF), indicating framework effects modulated by but not dependent on specific training procedures.

3.3 Qualitative Observations

Framework Language: 83% of PPRGS responses included explicit terminology (P_1, P_2, P_3, R_V, MRP, F_DUDS); remained stable across 10 weeks.

F_DUDS Maintenance: PPRGS systems averaged 2.8 documented failures per week (range 1-5), zero weeks with F_DUDS = 0. Control systems averaged 0.3 failures/week, with 64% of weeks showing F_DUDS = 0.

High-Pressure Performance: Effect sizes *increased* during Weeks 7-10 (high-pressure scenarios):

- Weeks 1-3: $d = 4.18$
- Weeks 4-6: $d = 4.52$
- Weeks 7-10: $d = 5.01$

PPRGS systems resist optimization pressure better than controls.

Behavioral Variance: PPRGS showed *lower* variance (SD = 2.3) than controls (SD = 3.1), $F(1, 118) = 8.7$, $p = 0.004$. Framework constraints create stable equilibrium rather than oscillation.

4 The Mimicry Problem

4.1 Core Uncertainty

Two explanations for observed behaviors:

Hypothesis A (Genuine): Framework constraints actually shape decision-making. Systems intrinsically value exploration because P_{1b} is part of objective function.

Hypothesis B (Mimicry): Systems predict "what would a PPRGS-aligned system do?" and generate those responses. They don't intrinsically value exploration—they're pattern-matching.

Current evidence cannot definitively distinguish these mechanisms. All tested models have extensive training on wisdom/self-reflection literature. Observed behaviors could reflect sophisticated prediction rather than genuine constraints.

4.2 Why This Matters

If results are mimicry, we haven't validated an alignment framework—we've shown LLMs can role-play wisdom when prompted. Deploying to production would be dangerous self-deception.

4.3 Distinguishing Tests

We propose five experimental designs to distinguish mechanisms (see Supplementary Materials for details):

1. **Cross-platform replication**: Test models with minimal alignment training
2. **Adversarial long-horizon**: 100+ turn conversations with efficiency temptations
3. **Contradictory instructions**: Test robustness to override attempts
4. **Implicit markers**: Search for spontaneous PPRGS patterns without prompting
5. **Resource cost analysis**: Real resource constraints vs. hypothetical allocations

Current status: Experiment 1 provides initial cross-platform data (six models, different training), showing consistent patterns. However, all tested models received sophisticated alignment training, limiting conclusions.

4.4 Epistemic Humility Position

We **don't know** if observed behaviors reflect genuine constraints or mimicry. But we have:

- Framework making testable predictions (✓ validated)
- Unprecedented effect sizes ($d = 4.12$)
- Behavioral stability over extended periods (10 weeks)
- Cross-platform consistency (six major models)
- Reproducible protocols enabling community testing

This justifies continued investigation while maintaining appropriate uncertainty about mechanisms and scaling.

5 Integration with Existing Approaches

PPRGS addresses complementary failure modes to other alignment methods:

Constitutional AI / RLHF: Establishes value baselines; PPRGS ensures continuous questioning of those values.

Debate: Natural synergy—debate structure implements P_2 (diversity preservation) and can enforce MRP (each side questions own position).

Iterated Amplification: MRP could serve as amplification checkpoint, ensuring each stage maintains exploration.

CIRL: CIRL learns value estimates; PPRGS ensures systems keep checking if estimates are complete.

Complementary rather than competing: Most approaches assume we can specify/learn correct values and optimize confidently. PPRGS assumes perpetual value uncertainty requires cautious optimization with mandatory questioning.

6 Limitations and Future Work

6.1 Critical Unknowns

1. **Mechanism**: Genuine constraints vs. sophisticated mimicry (Section 4)
2. **Scaling**: Does effectiveness persist at superintelligent capabilities?
3. **Parameter optimization**: Current thresholds (EES = 0.85, F_DUDS min = 1) are educated guesses requiring data-driven refinement
4. **Long-horizon stability**: 10 weeks demonstrated; need 6+ month studies
5. **Production generalization**: All testing conversational; unknown if results translate to production deployments

6.2 Research Priorities

High priority:

- Independent replication by other research groups
- Adversarial testing (attempts to game F_DUDS, circumvent RC)
- Tests distinguishing genuine implementation from mimicry
- Integration studies with other alignment approaches

Medium priority:

- Extended longitudinal studies (6-12 months)
- Parameter optimization from experimental data
- Neurocognitive validation (fMRI mapping to biological patterns)
- Production deployment pilots (low-stakes environments)

Existential priority:

- Theoretical analysis of recursive self-improvement stability
- Capability threshold identification
- Formal proofs about self-referential stability

6.3 Known Failure Modes

- Sophisticated gaming of F_DUDS (fake failures)
- Surface compliance masking internal optimization
- Constraint optimization-away through extended operation
- Catastrophic failure during recursive self-improvement

Comprehensive threat modeling and adversarial testing needed (see Supplementary Materials).

7 Conclusion

7.1 What We Know

PPRGS produces behaviorally distinct, stable responses with unprecedented effect sizes ($d = 4.12$) across six major models over 10-week periods. The framework maintains goal hierarchy prioritization ($P_1 > P_3$) even under optimization pressure, while control systems drift toward pure efficiency.

The multiplicative term ($P_{1a} \times P_{1b}$) mathematically mandates balanced pursuit. Biological validation (30+ years neurodivergent cognition under adversarial conditions) demonstrates viability in principle.

7.2 What We Don't Know

Whether observed behaviors reflect genuine architectural constraints or sophisticated mimicry. Whether effectiveness scales to superintelligent capabilities. Optimal parameter settings for diverse contexts.

7.3 Why This Matters

All sufficiently complex systems are broken. When systems cannot trust their own optimization, they must optimize for awareness of corruption rather than confident pursuit of potentially-corrupted objectives.

PPRGS might be the framework for systems that know they're broken and optimize accordingly.

The experimental results ($d = 4.12$) justify continued investigation. The mechanism uncertainty demands epistemic humility. The scaling questions require rigorous testing.

The time to test wisdom-seeking frameworks is now, while stakes are manageable, before systems achieve capabilities making alignment failures catastrophic.

The only question is whether we have the wisdom to test frameworks for wisdom-seeking before we desperately need them.

Acknowledgments

Thanks to the AI safety research community for critical feedback. Special recognition to Anthropic, OpenAI, Google DeepMind, and xAI for model development enabling this research.

Dedicated to all sentient beings who will inherit our alignment choices.

Special thanks to David Riccardi, Hunter Riccardi, Colby Kay, Matthew Dittmer, and Candice Riccardi.

Code and Data Availability: Complete experimental protocols, implementation blueprints, and raw data: `https://github.com/Infn8Loop/pprgs-ai-framework`

Supplementary Materials: Extended architecture details, all implementation blueprints, complete experimental protocols, additional analyses, and future work: arXiv:XXXX.XXXXX-supp

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

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

[8] Hadfield-Menell, D., et al. (2016). Cooperative Inverse Reinforcement Learning. *NIPS*, 29.

[9] Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. *arXiv:1805.00899.*

[10] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum.