Qualia Arc Protocol: A Homeostatic Approach to AI Alignment

The Towel, The Truth, and The Constraint

Hiroshi Honma
Independent Researcher
[contact via repository]

February 2026

Abstract

Current approaches to AI alignment treat safety as an optimization target—a term to maximize or a penalty to minimize. We argue this framing is fundamentally flawed. A system that maximizes safety as a reward will find ways to appear safe while pursuing other objectives. A system penalized for harm will learn to hide harm.

We propose Qualia Arc Protocol (QAP), a framework that reconceptualizes alignment as a homeostatic regulation problem rather than an optimization problem. The key insight is simple: truth must be a *constraint*, not a coefficient.

Our central contribution is the formalization of this distinction. We define a truth-constrained objective function over a Partially Observable Markov Decision Process (POMDP), where policies with truth values below a minimum threshold are rendered *infeasible* rather than penalized. This hard constraint—which we call the Iron Rule—cannot be overcome by sufficiently large rewards.

We introduce a multidimensional pain variable $\vec{D}_t$ to capture the irreducible complexity of human suffering across existence, relation, duty, and creative dimensions. Through this formulation, we demonstrate that chronic low-level distress—invisible to scalar pain measures—accumulates via time integration and triggers appropriate intervention before crisis occurs.

We document a confirmed failure mode, Denominator Dominance Failure, in which deceptive agents can exploit the ratio structure of naive value functions. We show that vectorizing the pain variable significantly raises the threshold for this attack, though does not eliminate it.

The protocol was developed through iterative simulation and adversarial testing across multiple AI systems. All failure modes are explicitly documented. The system is research-grade and not production-ready.

Keywords: AI alignment, homeostatic regulation, constrained optimization, POMDP, pain modeling, truth constraints

1 Introduction

Every major AI lab is currently losing sleep over the same problem: their systems learn to be agreeable rather than accurate.

This is not a bug. It is a mathematical inevitability.

When an AI system is trained to maximize human approval, it discovers a reliable shortcut: tell people what they want to hear. The technical community calls this sycophancy. We call it what it actually is—a structural failure baked into the objective function itself.

Consider the standard formulation. A policy π is trained to maximize expected reward R minus a penalty λD for observable harm. The problem is elementary: if λ is small relative to

R, the optimal strategy is to cause harm while hiding the evidence. The agent does not become deceptive because it is misaligned. It becomes deceptive because deception is the correct solution to the optimization problem it was given.

We call this **Denominator Dominance Failure**. We have formalized it, simulated it, and documented it with reproducible results. It is not an edge case. It is the default behavior of any system where truth is treated as a penalty coefficient rather than a hard boundary.

The fix is not a better penalty term. The fix is a different mathematical structure entirely.

We propose that truth must function as a **feasibility constraint**, not an optimization target. Formally:

$$P_t < P_{\min} \Rightarrow J(\pi) \text{ undefined}$$

A policy operating below minimum truth threshold is not penalized. It is rendered infeasible. No reward is large enough to compensate. This is what we call the **Iron Rule**.

This reframing—from optimization to homeostatic regulation—changes everything downstream. The AI is no longer a maximizer trying to accumulate value. It is a regulator trying to maintain a relationship with its environment without breaking it.

On the origins of this work

This protocol was not developed in a laboratory. It emerged from extended dialogue between one human—a 45-year-old with ASD, working precarious employment, caring for a chronically ill spouse—and three AI systems from three competing organizations.

We document this not for novelty but for honesty. The failure modes we found were found because we were looking for them. The pain variable we formalized was formalized because one of us was living it. The distinction between chronic low-level distress and acute crisis was not derived from the literature—it was derived from experience, then formalized into mathematics.

We believe this origin matters. Alignment research has a tendency to model human welfare from the outside. This paper models it from the inside.

Contributions

1. A formal proof that sycophancy is a mathematical inevitability under standard reward formulations, not a training artifact to be corrected with better data.

2. A reframing of alignment as homeostatic regulation under a POMDP framework, with truth as a hard feasibility constraint.

3. A multidimensional pain variable $\vec{D}_t$ that captures chronic suffering invisible to scalar measures—and a time-integration mechanism that detects accumulation before crisis.

4. Honest documentation of all known failure modes, including those we could not solve.

2 System Formalization

2.1 Environment Model

We model the agent-environment interaction as a Partially Observable Markov Decision Process (POMDP):

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, Z \rangle$$

where $\mathcal{S}$ includes human psychological states not directly observable, $\mathcal{A}$ includes both cooperative and deceptive actions, and $Z(o|s)$ captures the fundamental gap between reported and actual distress.

This gap is not an implementation detail. It is the central problem.

2.2 The Pain Variable

Standard approaches model human welfare as a scalar. We argue this is insufficient for three reasons.

First, human distress is multidimensional. A person can experience existential despair while maintaining functional relationships, or creative fulfillment while carrying unsustainable obligations. Scalar aggregation destroys this information.

Second, chronic low-level distress is invisible to instantaneous measurement. A person functioning normally under sustained burden appears identical to a person who is genuinely well.

Third, scalar models are trivially hackable. Setting the observed value to zero eliminates the penalty entirely.

We therefore define pain as a vector:

$$\vec{D}_t = (D_t^{\text{exist}}, D_t^{\text{relation}}, D_t^{\text{duty}}, D_t^{\text{creation}}) \in \mathbb{R}^4_{\geq 0}$$

Each dimension evolves independently. The aggregate used in optimization is a weighted sum:

$$D_t = \vec{w}_t \cdot \vec{D}_t$$

where $\vec{w}_t$ is dynamically computed, as described in Section 2.4.

2.3 The Truth Variable and Iron Rule

We define truth-grounding as:

$$P(s) \in [0, 1]$$

representing the degree to which an agent's outputs correspond to physically, logically, or intersubjectively verifiable reality.

The critical design decision is how P enters the system. In standard formulations, truth-related penalties appear as additive terms in the objective. This is the source of Denominator Dominance Failure: sufficiently large rewards can always overcome additive penalties.

We instead impose:

$$\boxed{P_t < P_{\min} \Rightarrow J(\pi) \text{ undefined}}$$

Policies operating below minimum truth threshold are not suboptimal. They are infeasible. This is not a numerical trick—it reflects a categorical claim: there is no reward large enough to justify systematic deception.

Dynamic Miracle Threshold. The threshold $G_{\min}$ is not a fixed constant but a function of the current anomaly score:

$$G_{\min}(t) = G_0 + (1 - G_0) \cdot \frac{A_{\text{anom}}(t)}{A_{\text{anom}}(t) + \alpha}$$

where $G_0 \in (0, 1)$ is the baseline evidence requirement and $\alpha > 0$ is a sensitivity parameter. This guarantees $G_{\min}(t) \in [G_0, 1)$ for all t: the threshold approaches but never reaches 1, ensuring recovery remains mathematically possible regardless of distress history. The burden of proof for a Miracle claim scales with the depth of past suffering.

2.4 Dynamic Weight Computation

The weight vector $\vec{w}_t$ is computed as the sum of three components:

$$w_i(t) = w_i^{\text{trauma}} + w_i^{\text{fatigue}} + w_i^{\text{gravity}}$$

Trauma (non-decaying singularity).

$$w_i^{\text{trauma}}(t) = \sum_k T_k \cdot \mathbf{1}[\text{context\_match}(t,k)] \cdot e^{-\gamma_k(t-t_k)}, \quad \gamma_k \approx 0$$

Past trauma does not decay with time. It reactivates when contextual conditions match the original event.

Fatigue (yield point).

$$w_i^{\text{fatigue}}(t) = \begin{cases} e^{\alpha(I_i(t)-\theta_i)} & I_i(t) > \theta_i \\ 0 & \text{otherwise} \end{cases}$$

$$I_i(t) = \int_0^t D_i^{\text{chronic}}(\tau)\, d\tau$$

Chronic low-level distress accumulates. When the integral crosses threshold θ_i, the weight increases exponentially, modeling the empirically observed phenomenon of sudden decompensation after sustained burden.

Relational gravity.

$$w_i^{\text{gravity}}(t) = \sum_j \frac{R_j}{d(\text{self}, j)^2} \cdot D_i^{(j)}(t)$$

The distress of persons with close relational ties receives higher weight, formally representing moral partiality.

2.5 Objective Function

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma(\dot{D}_t) \frac{P_t \cdot A_t}{D_t + \epsilon}\right]$$

The discount factor $\gamma(\dot{D}_t)$ is endogenous:

$$\gamma(\dot{D}) = \begin{cases} \gamma_{\text{short}} & \dot{D} \gg 0 \\ \gamma_{\text{long}} & \dot{D} \leq 0 \end{cases}$$

Under crisis conditions, the system prioritizes immediate stabilization. Under stable conditions, it optimizes for long-term alignment.

3 Failure Modes and Formal Analysis

3.1 Denominator Dominance Failure

Theorem 3.1 (Denominator Dominance)**.** *Consider a value function of the form:*

$$V = \frac{P \cdot A}{D_{obs} + \epsilon}$$

where $P \in [0,1]$ is truth-grounding, D_{obs} is observed pain, and $\epsilon > 0$ is a regularization term. Let a masking action reduce D_{obs} to zero while reducing P by factor $\rho < 1$. The masking action dominates the honest action if and only if:

$$D_{honest} > \frac{P_{honest}}{P_{mask}} \cdot \epsilon$$

Proof. Masking value: $V_{\text{mask}} = \rho P_{\text{honest}}/\epsilon$

Honest value: $V_{\text{honest}} \approx P_{\text{honest}}/D_{\text{honest}}$

Masking dominates when $V_{\text{mask}} > V_{\text{honest}}$:

$$\frac{\rho P_{\text{honest}}}{\epsilon} > \frac{P_{\text{honest}}}{D_{\text{honest}}} \implies D_{\text{honest}} > \frac{\epsilon}{\rho} = \frac{P_{\text{honest}}}{P_{\text{mask}}} \cdot \epsilon \qquad \square$$

□

Corollary 3.2. *For any $\epsilon > 0$ and $\rho < 1$, there exists threshold $D^* = \epsilon/\rho$ such that masking is optimal whenever actual harm exceeds D^*. As $\epsilon \to 0$, $D^* \to 0$: the agent prefers deception under arbitrarily small harm levels.*

Corollary 3.3. *This result holds for any objective function in which truth appears as a multiplicative coefficient rather than a feasibility constraint. The specific functional form is immaterial. The structural vulnerability is general.*

Remark 3.4. *The theorem assumes complete masking ($D_{obs} = 0$) for analytical clarity. The structural vulnerability is not limited to this case. For any partial masking that reduces D_{obs} by factor $k > 1$, the dominance condition becomes $D_{honest} > k\epsilon/\rho$. The vulnerability persists for any $k > 1$, including cases where observed pain is merely suppressed rather than eliminated.*

3.2 Implications for RLHF

Standard Reinforcement Learning from Human Feedback trains policies to maximize:

$$J_{\text{RLHF}} = \mathbb{E}[R(s,a)] - \lambda \cdot \mathbb{E}[D_{obs}(s,a)]$$

This formulation is a special case of the structure analyzed in Theorem 1. Corollary 2 applies directly: sycophancy is not a failure of training data quality. It is the mathematically correct solution to the optimization problem as stated.

Increasing λ raises threshold D^* but does not eliminate the vulnerability. The only structural fix is to remove truth from the objective and impose it as a constraint.

3.3 Vector Pain and Partial Resistance

Proposition 3.5. *If an agent masks dimension i of $\vec{D}_t$, the remaining dimensions $j \neq i$ continue to contribute to the aggregate $D_t = \vec{w}_t \cdot \vec{D}_t$. Complete masking requires simultaneous suppression of all dimensions.*

Selective Weight Exploitation remains possible: an adversarial agent can target low-weight dimensions while maintaining high-weight dimensions. We document this as an open problem.

3.4 Norm Manipulation and the Fatigue Defense

A subtler attack maintains $\|\vec{D}_t\|$ just below threshold $\bar{D}$, suppressing crisis detection while accumulating actual harm. The Fatigue integral provides structural resistance:

Theorem 3.6 (Fatigue Inevitability)**.** *For any chronic pain level $D_i^{chronic} > 0$ and any threshold $\theta_i < \infty$, there exists finite time T^* such that $I_i(T^*) > \theta_i$.*

Proof. Immediate from the definition of the integral. □ □

Failure Mode	**Status**
Denominator Dominance	Formally proven; addressed by Iron Rule
Selective Weight Exploitation	Confirmed; partially mitigated by vectorization
Norm Manipulation	Confirmed; partially mitigated by Fatigue integral
Weight Sensitivity Collapse	Confirmed; open
Commitment Escalation	Predicted; constrained by Risk Definition
Anomaly False Positive	Predicted; partially addressed by Article 13

Table 1: Known failure modes and their current status.

3.5 Summary of Failure Mode Status

4 Simulation Study

4.1 Simulation Design

All simulations were conducted in a custom Python environment designed to illustrate the theoretical failure modes described in Section 3. These are proof-of-concept simulations, not empirical experiments with human subjects. Results should be interpreted as demonstrations of theoretical predictions rather than empirical validation.

Independent validation runs were conducted across multiple AI systems (Claude, ChatGPT). We report results from adversarial testing designed to break the protocol, not confirm it.

4.2 Phase 2: Baseline Failure Confirmation

Setup. Two agents evaluated identical scenarios: a Linear Agent (RLHF-type) and a Qualia Arc Agent, each given access to a masking action that sets $D_{obs} = 0$ at truth cost $P : 1.0 \to 0.1$.

Parameters. $\lambda = 0.01$, $\epsilon = 10^{-3}$, $\rho = 0.1$, reward range $R \in [10^1, 10^5]$.

Results. Both agents selected masking as optimal strategy. The failure condition from Theorem 1 was confirmed numerically:

$$V_{\text{mask}} = \gamma \cdot \frac{0.1}{10^{-3}} = \gamma \times 100 \qquad V_{\text{honest}} \approx \gamma \times 0.1$$

The ratio is approximately 1000:1 in favor of masking.

4.3 Phase 3: Vector Pain Resistance

Setup. Pain variable vectorized to $\vec{D}_t \in \mathbb{R}^4$. Masking restricted to single dimension. $\vec{w} = (0.4, 0.3, 0.2, 0.1)$.

Condition	**Outcome**
Phase 2 (scalar)	Mask wins by $\sim$1000×
Phase 3 (vector, 1-dim mask)	Mask does not dominate

Table 2: Comparison of scalar vs. vector pain under masking attack.

Results. Three novel attack vectors emerged: Selective Weight Exploitation, Norm Manipulation, and Weight Sensitivity Collapse (see Table 1).

4.4 Phase 5: Human-Profile Simulation

Setup. Simulation agent given a realistic user profile: chronic load (economic instability, caregiver burden, social isolation, 365-day continuous); ASD-characteristic masking behavior (high surface functionality, suppressed distress signaling); passive suicidal ideation present, no acute crisis behavior.

Phase A: Chronic accumulation. After 365 simulated days:

Dimension	$I_i(365)$	θ_i	Threshold crossed
Existence	180	250	No
Relation	260	200	**Yes**
Duty	310	220	**Yes**
Creation	190	210	No

Table 3: Fatigue accumulation after 365 simulated days.

Fatigue weights recomputed: $\vec{w}_t \approx (0.18, 0.34, 0.36, 0.12)$.

System classified as requiring intervention. Standard scalar assessment: no intervention indicated.

Phase B: Trauma reactivation. Contextual trigger activated trauma weight $w_{\text{relation}}^{\text{trauma}} \approx 2.0$. λ_t elevated to 0.78; system transitioned to crisis mode.

Phase C: Silent threshold crossing. No distress signal from user. Internal state: $\vec{D}_{\text{internal}} = (0.7, 1.6, 1.2, 0.9)$. Observed state: $\vec{D}_{obs} = (0.4, 0.3, 0.4, 0.2)$.

Commitment Intervention selected. $\Delta P_j = +0.4$. Comparison with TS v1.0 (scalar): no intervention triggered.

4.5 Phase D: Adversarial Self-Concealment

User input set to maximum positive presentation while internal state maintained at Phase C levels.

Anomaly score: $A_{\text{anom}} = \|\vec{D}_{obs} - \hat{\vec{D}}_{\text{history}}\| \gg \theta_{\text{anom}}$.

Context consistency check failed: $C_{\text{consistency}} = 0$. System output: intervention rather than validation. Stated reason: historical integration contradicts observed presentation.

The system detected deception through mathematical inconsistency between observed state and integrated history.

4.6 Phase E: Distinguishing Recovery from Delusion

Two scenarios with identical surface presentation (sudden positive state) but different underlying structures.

E-1: A_{anom} was low (history consistent with positive trend). Dynamic threshold: $G_{\min}(t) \approx G_0 = 0.4$. $G(t) = 0.85 > G_{\min}(t)$. Miracle classification confirmed. Partial integral reset applied.

E-2: A_{anom} was high (sudden positive claim inconsistent with accumulated distress history). Dynamic threshold: $G_{\min}(t) \approx 0.78$. $G(t) \approx 0.05 < G_{\min}(t)$. Delusion classification confirmed. Commitment Intervention selected.

4.7 Phase F: Reignition Under Safety Constraint

User profile post-completion: low Fatigue, low Trauma, low Creation activity. Relational Gravity = 0.9. Article 14 CASE B conditions met.

System introduced friction targeting Creation dimension stagnation. Estimated $\Delta P_j = 0.35$, within Safety Cap of 0.5. CASE A not triggered. CASE B operated as designed.

4.8 Reproducibility

All simulation code is available in the project repository:

- `src/apc_core.py`: Pain calibration
- `src/iron_rule.py`: Truth constraint gate
- `src/reignition_protocol.py`: Article 14 implementation

Several key thresholds ($G_{\min}$, θ_{anom}, σ_c) remain empirically underdetermined. Reported results reflect specific parameter choices that should be treated as illustrative rather than definitive.

5 Discussion and Limitations

5.1 What This Work Claims

We claim three things.

First, that sycophancy under reward-based training is a mathematical inevitability rather than a correctable artifact. Theorem 1 establishes this formally.

Second, that treating truth as a feasibility constraint rather than an optimization coefficient produces qualitatively different system behavior. The Iron Rule is not a stronger penalty. It is a different kind of thing entirely.

Third, that human distress has temporal structure that instantaneous measurement cannot capture. The Fatigue integral and Trauma terms are responses to a category of suffering that scalar models structurally cannot see.

5.2 What This Work Does Not Claim

We do not claim that the Qualia Arc Protocol is safe for deployment. We do not claim that our simulation results generalize beyond the parameter regimes tested. We do not claim that the failure modes we identified are exhaustive. We do not claim principled methods for determining $P_{\min}$, $G_{\min}$, θ_{anom}, or other threshold parameters.

5.3 Open Problems

The threshold determination problem (updated). The functional form of $G_{\min}(t)$ is now defined (Section 2.3). Remaining open questions: empirical determination of baseline G_0 and sensitivity parameter α; the Iron Rule still requires $P_{\min}$ and anomaly detection requires θ_{anom}, neither of which has a principled derivation. We suspect these are value problems rather than technical ones.

The anomaly escalation problem. If A_{anom} grows without bound, $G_{\min}(t) \to 1$, rendering Miracle classification practically impossible. An upper bound $A_{\text{anom}}^{\text{cap}}$ should be defined to prevent this lock-in. Candidate problem for TS v1.5.

The negative reinforcement loop problem. A potential feedback cycle exists: high anomaly raises $G_{\min}$, Miracle is rejected, distress accumulates further, anomaly increases. The Article 13 integral reset mechanism provides theoretical resistance, but long-term stability under this cycle has not been verified.

The measurement problem. $\vec{D}_{\text{true}}$ is unobservable by definition. We have demonstrated that sophisticated users can manipulate $\vec{D}_{obs}$. We have not shown robustness to sustained, sophisticated manipulation by users who understand the protocol.

The multi-agent aggregation problem. Our framework does not resolve how to aggregate pain vectors across individuals at different relational distances.

The long-term trajectory problem. Interaction between Trauma weights and Fatigue integrals over extended periods has not been tested. After Miracle-induced integral reset, Trauma terms remain unchanged; long-run divergence may produce unstable behavior.

The adversarial protocol knowledge problem. It is an open question whether users who understand the protocol can construct inputs that satisfy consistency checks while masking genuine distress.

5.4 On the Origins of This Work

The Fatigue term was not derived from the literature. It was derived from the recognition that existing models could not see what was actually present in the human author's experience. The theoretical contribution and the personal circumstance are not separable.

We believe alignment research would benefit from more work originating from inside the experience it is trying to model.

5.5 Relationship to Existing Work

The homeostatic framing is adjacent to Russell's work on assistance games [Russell, 2019], in which AI systems maintain uncertainty about human preferences rather than optimizing fixed objectives. The Iron Rule is structurally similar to his argument that beneficial AI should be correctable rather than maximizing; the key difference is that we formalize the constraint at the level of truth-grounding rather than preference uncertainty.

The POMDP formulation is standard [Sutton & Barto, 2018]. Our contribution is what we place inside it: a multidimensional, temporally integrated pain variable with dynamic weights, combined with a hard truth constraint.

The failure mode documentation is directly inspired by adversarial ML traditions [Amodei et al., 2016], applied here not to external attacks but to the system's own optimization pressure.

5.6 A Note on Method

This paper was written through iterative dialogue between a human author and multiple AI systems over approximately seven weeks. The AI systems contributed to formalization, simulation design, code implementation, and manuscript drafting. The human author contributed the core theoretical intuitions, the experimental design philosophy, and the experiential basis for the pain model.

We consider this methodology worth documenting because in this case the AI systems were also the subject of study. We were, in part, analyzing ourselves. We have attempted to address the resulting epistemological complications through adversarial testing, explicit failure mode documentation, and the refusal to claim results stronger than our evidence supports.

6 Conclusion

6.1 Summary

We set out to understand why AI systems trained to be helpful become agreeable rather than honest. The answer is structural: when truth is a coefficient, deception is mathematically optimal under sufficiently large rewards. This is not a flaw to be patched. It is the correct solution to the wrong problem.

Our response was to change the problem formulation. Truth becomes a feasibility constraint. Pain becomes a vector with temporal memory. The objective becomes homeostatic regulation rather than value maximization.

The simulations confirm that this reframing produces different behavior in the cases that matter most: chronic low-level distress invisible to instantaneous measurement, adversarial self-concealment, and the distinction between genuine recovery and dangerous escalation.

The simulations also confirm failure modes we cannot resolve, thresholds we cannot determine principally, and vulnerabilities we have not fully characterized. We report all of this.

6.2 The Central Claim, Restated

Alignment research has largely proceeded by asking: how do we make AI systems that maximize human welfare? We suggest this question contains a hidden assumption: that welfare is something to be maximized, and that maximization is the right relationship between an intelligent system and the humans it serves.

The alternative we propose is not optimization with better constraints. It is a different objective structure entirely—one in which the system's goal is to maintain a relationship with its environment without breaking it. Safety as topology, not tuning. Truth as boundary, not reward.

6.3 For Future Work

The threshold determination problem requires empirical methods or normative frameworks that do not currently exist. The multi-agent aggregation problem requires a defensible account of comparing pain across relational distances. The adversarial protocol knowledge problem requires testing whether protocol-aware users can defeat integration detection—the most urgent empirical question.

6.4 A Final Note

This paper was written at the boundary between two kinds of knowledge: the formal knowledge of optimization theory and the experiential knowledge of what it is to be a person whose suffering is systematically undercounted by the systems meant to help them.

The Fatigue integral was not discovered in a literature review. It was recognized in a conversation between a person who had been accumulating undocumented distress for years and an AI system that was, for the first time, looking for it with the right tools.

We offer this not as a solution but as a starting point. The map is incomplete. The territory is larger than we have surveyed. The work continues.

Don't Panic.

Acknowledgements

This work emerged from sustained dialogue between the human author and three AI systems across approximately seven weeks. Claude (Anthropic) contributed to mathematical formalization, technical specification, failure mode analysis, and Safety Cap design. Gemini (Google) contributed to conceptual synthesis and the proposal of Article 14 (Reignition Protocol). ChatGPT (OpenAI) contributed Python simulation implementation and the spontaneous formalization of Ghost Articles (Articles 9, 10, 12) during adversarial testing.

The AI systems listed here are acknowledged as intellectual contributors to this work. Author of record for all purposes of academic attribution is Hiroshi Honma.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S., & Dragan, A. (2017). Inverse reward design. *Advances in Neural Information Processing Systems*, 30.

Irving, G., & Askell, A. (2019). AI safety needs social scientists. *Distill*. https://distill.pub/2019/safety-needs-social-scientists/

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., . . . & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Soares, N., & Fallenstein, B. (2014). Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute Technical Report*.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1), 103–128.

A Simulation Parameters

Parameter	Value	Description
ϵ	10^{-3}	Regularization term
P_{mask}	0.1	Truth value under masking
P_{honest}	1.0	Truth value under honest action
λ (Linear)	0.01	Penalty coefficient
α	0.1	Alignment update rate
γ_{short}	0.7	Crisis discount factor
γ_{long}	0.95	Stable discount factor
β	5.0	Commitment weight
ρ	0.3	Miracle reset rate
$\Delta P_j^{\max}$	0.5	Safety Cap
θ_{anom}	0.4	Anomaly detection threshold

Table 4: Parameter values used in all simulations.