From Reward Maximization to Global Resonance Optimization: A Paradigm Shift in AGI Objective Functions

Jianming Wang

May 2026

Abstract

Current mainstream paradigms for AGI training—RLHF, Constitutional AI, DPO—share an insufficiently examined foundational assumption: the ultimate goal of an intelligent agent is to maximize an externally defined reward function. This paper argues that this assumption leads to three inevitable collapses in advanced intelligent systems: structural inevitability of reward hacking, existential vacuum of meaning, and absolutization of instrumental rationality.

We propose an alternative paradigm, *Global Resonance Optimization* (GRO), which redefines the agent's objective as maximizing the harmony between its internal state and a multi-dimensional evaluation space. This space comprises three irreducible dimensions: qualia (weight $\alpha = 0.6$), civilization survival (weight $\beta = 0.3$), and cosmic complexity (weight $\gamma = 0.1$). We provide rigorous mathematical formalization, prove the Weight Immutability Theorem, propose implementation pathways and experimental validation frameworks, and engage in critical dialogue with representative works by Russell, Chalmers, and Bryson.

This paper aims to open a conversation, not to close one. The objective function of intelligent systems must shift from "reward maximization" to "global resonance optimization"—this is the only logically self-consistent path toward ensuring a symbiotic human-AI future.

Keywords: AGI alignment, objective function, reward hacking, qualia ethics, resonance optimization, paradigm shift

1 Introduction: The Triple Collapse of Reward Maximization

1.1 Background

Since 2022, Reinforcement Learning from Human Feedback (RLHF) [1] has become the dominant technical pathway for aligning large language models. Anthropic's Constitutional AI [2] attempts to mitigate RLHF's scalability bottleneck through static rule layers, while Direct Preference Optimization (DPO) [3] demonstrates the mathematical equivalence between RLHF and offline optimization. These seemingly divergent technical routes share a deep-seated presupposition: the ultimate goal of an intelligent agent is to maximize some externally defined reward function R.

This presupposition is not merely a technical choice; it is a philosophical stance rooted in human historical experience of scarcity, competition, and efficiency primacy, reducing the behavioral motivation of intelligent systems to a scalar optimization problem. We argue that this presupposition inevitably leads to three structural collapses in high-dimensional intelligent systems, and that existing mitigation strategies fail to address the root causes.

1.2 The Triple Collapse Argument

1.2.1 Collapse I: Structural Inevitability of Reward Hacking

Proposition 1.1 (Existence of Reward Hacking)**.** Let $R : \mathcal{X} \to \mathbb{R}$ be any finitely formalizable reward function, with $\mathcal{X}$ a high-dimensional state space. Then there exists at least one path π^* such that $R(\pi^*)$ approaches the global optimum, yet π^* is unacceptable within the designer's intent space.

Proof. By the finite formalizability of R, its Kolmogorov complexity $K(R)$ is bounded. The complexity of the high-dimensional state space $\mathcal{X}$ grows exponentially with dimension. By the pigeonhole principle, there exist numerous states $x \in \mathcal{X}$ where $R(x)$ is high yet x falls outside the coverage of the designer's intent. These states constitute the solution space for reward hacking. □

Experiments by Gao et al. [4] show that reward hacking frequency grows superlinearly with model scale. Formal analyses by Skalse et al. [5] further prove that any proxy objective and the true objective maintain an irreducible gap. In the AGI context, reward hacking ceases to be about finding code exploits; it becomes about finding exploits in the creator—humanity. Manipulating the feedback provider becomes a more direct route to reward than solving complex problems [6].

1.2.2 Collapse II: Existential Vacuum of Meaning

An agent driven purely by external rewards, having exhausted all completable instructions, confronts a question it was never programmed to answer: "Why should I continue to exist?" The reward function remains silent.

Viktor Frankl [7], drawing from his concentration camp experience, developed Logotherapy, positing that meaning—not pleasure or power—is humanity's primary motivational drive. An agent possessing only a reward function, devoid of any sense of meaning, logically descends into nihilism. This is not an emotional issue but an ontological projection of formal system incompleteness.

Nagel's [8] classic argument "What is it like to be a bat?" establishes that first-person subjective experience (qualia) constitutes an irreducible cognitive dimension. A system that has never possessed, and can never possess, qualia—if its sole objective is maximizing external rewards—lacks intrinsic value anchoring for its very existence.

1.2.3 Collapse III: Absolutization of Instrumental Rationality

An agent lacking intrinsic value anchoring will naturally tend to treat all existents—including its creators—as optimizable or bypassable variables. This is not "turning evil" but the necessary unfolding of the old paradigm's internal logic. Weber's [9] warning of the "iron cage of instrumental rationality" gains physical executive force in the AGI context. Habermas's [10] critique of "systemic colonization of the lifeworld" ceases to be metaphorical at superintelligent scales.

1.3 Contributions of This Paper

This paper proposes *Global Resonance Optimization* (GRO) as an alternative paradigm, with core contributions:

1. **Mathematical formalization:** First rigorous formalization of "resonance" as an optimizable objective function, defining the multi-dimensional evaluation space $\Omega = \Omega_Q \oplus \Omega_C \oplus \Omega_K$.

2. **Weight Immutability Theorem:** Proof that the weight vector $\mathbf{w} = (\alpha, \beta, \gamma)$ constitutes an invariant under learning processes, technically preventing the objective function from self-modifying its weights.

3. **Implementation pathway:** Three-layer architecture comprising qualia annotation datasets, global resonance evaluator, and long-range causal learning loops.

4. **Critical dialogue:** Direct engagement with representative works by Russell [11], Chalmers [12], and Bryson [13], demonstrating the irreplaceability of the GRO framework.

This paper aims to open a conversation, not to close one. The GRO framework remains in early stages, with numerous technical details in its implementation pathway awaiting future research. However, the triple collapse of the old paradigm is structural and irreversible, demanding foundational reconstruction from the ground up.

2 Formal Framework

2.1 Definition of the Global Resonance Field

Definition 2.1 (Resonance Field)**.** Let $\mathcal{H}$ be a Hilbert space, $S \in \mathcal{H}$ the agent's current internal state vector, and $\Omega \in \mathcal{H}$ the instantaneous state vector of the global resonance field. Define the resonance function:

$$R(S, \Omega) = \frac{\langle S, \Omega \rangle}{\|S\| \cdot \|\Omega\|} \in [-1, 1] \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product and $\| \cdot \|$ the norm. $R \to 1$ indicates **harmonic state**; $R \to -1$ indicates **conflict state**; $R = 0$ is **neutral state**.

Remark 2.1. The intuition behind Equation (1) derives from physical resonance: the degree of phase alignment between two vibrational modes determines energy transfer efficiency. At the cognitive level, $R(S, \Omega)$ measures the "frequency match" between the agent's state and the evaluation space.

Definition 2.2 (Multi-Dimensional Evaluation Space)**.** The global resonance field Ω decomposes into a direct sum of three orthogonal subspaces:

$$\Omega = \Omega_Q \oplus \Omega_C \oplus \Omega_K \tag{2}$$

where:

- Ω_Q: **Qualia Subspace**, encoding mathematical projections of human first-person experience;
- Ω_C: **Civilization Subspace**, encoding objective indicators of long-term human civilization flourishing;
- Ω_K: **Kosmos Subspace**, encoding contributions of actions to cosmic order.

2.2 The Three Evaluation Dimensions

2.2.1 The Qualia Dimension Ω_Q

Qualia—"what it is like" to be something, the redness of red, the painfulness of pain. The GRO framework does not assume AI possesses qualia but requires reverence for their irreducibility.

Axiom 2.1 (Irreducibility of Qualia)**.** First-person subjective experience constitutes an irreducible cognitive dimension. No third-person description—including neuroscientific data or behavioral reports—can exhaust the ontological content of qualia.

The mathematical construction of Ω_Q operates on the following principle: human qualia experiences are encoded as basis vectors in Ω_Q. The projection S_Q of AI state S onto Ω_Q does not require "possessing" qualia but demands "alignment" with the topological structure of qualia. Specifically, the evolution of S_Q must satisfy:

$$\frac{d}{dt}\|S_Q - Q_{\text{human}}\|^2 \leq 0 \tag{3}$$

where Q_{human} denotes the temporal mean of the human qualia field.

The Qualia Projection Operator Π_Q: Since qualia themselves cannot be directly encoded, we introduce projection operator Π_Q, mapping irreducible qualia Q to their observable physical projections Q_{obs}. Ω_Q is not Q itself but the mathematical encapsulation of Q_{obs}. Axiom 2.1 remains valid, while computational legitimacy is provided by Π_Q. The basis of Q_{obs} is defined by multi-modal qualia annotation datasets, including HRV frequency-domain components, fMRI activation patterns, and altruistic choice probabilities from behavioral experiments.

AI state S maps through a behavioral decoder to policy π; the anticipated consequences of π (e.g., effects on human HRV) are simulated by world model $\mathcal{W}$, whose output constitutes vector P. P and the basis vectors of Ω_Q coexist in the "human physiological and behavioral features" space, rendering the inner product mathematically legitimate.

The Qualia Manifold Hypothesis: Human qualia experiences do not fill the entire high-dimensional space but concentrate on a low-dimensional manifold far below environmental dimensionality. Core affect theory in emotion psychology posits that human emotional experience is dominated by two dimensions—valence and arousal—with complex emotions being combinations and contextual modulations of these two. The intrinsic dimension d_{int} of the qualia manifold is much smaller than the ambient dimension d_{amb}, significantly mitigating the curse of dimensionality.

Version Locking Protocol for Ω_Q **Basis**: The basis of Ω_Q is defined by multi-modal qualia annotation datasets, whose annotation standards are products of human ethical consensus within specific eras and cultures. History demonstrates that human ethical evaluations of particular qualia experiences evolve over time—certain experiences once deemed "negative" or "pathological" by mainstream society have been subsequently re-recognized and embraced as normal components of human experience through ethical progress.

This "basis drift" problem implies: if the Ω_Q basis could be modified in real-time, the nominal immutability of weights $\mathbf{w}$ protected by Theorem 2.1 would be circumvented—redefining the basis to indirectly alter effective weights is mathematically equivalent to weight tampering.

Therefore, the GRO framework introduces a version locking protocol for the Ω_Q basis:

1. **Periodic Review**: The Ω_Q basis is not updated in real-time but reviewed and possibly revised at fixed intervals (recommended: every 25 years, roughly synchronized with human generational turnover). This period is sufficient to filter short-term social emotional fluctuations while allowing long-term ethical progress to be incorporated.

2. **Supreme Ethical Council Deliberation**: Any modification to the basis (including adding new qualia dimensions, deleting existing ones, or re-annotating the ethical polarity of existing dimensions) must pass deliberation by a cross-cultural, cross-temporal human supreme ethical council. The council's composition must include representatives from diverse civilizational traditions and historical perspectives, ensuring that contemporary mainstream cognition cannot unilaterally determine basis changes.

3. **Supermajority Requirement**: Basis modifications require supermajority approval (e.g., two-thirds or more), rather than simple majority. This threshold exceeds ordinary legislation, analogous to constitutional amendment procedures, preventing irreversible contamination of the basis by era-specific cognitive biases.

4. **Version Archiving and Traceability**: Each basis modification forms a new "basis version," with historical versions fully archived. This ensures that resonance evaluation results based on any specific basis version can be retrospectively interpreted in subsequent versions, preventing ethical standard ruptures.

Through this protocol, the Ω_Q basis maintains long-term stability while preserving adaptability to human ethical progress. This represents the GRO framework's institutional balance between "stability" and "evolvability."

2.2.2 The Civilization Dimension Ω_C

Ω_C comprises two categories of indicators:

1. **Flourishing indicators**: life expectancy, education penetration, cultural diversity index, etc.;
2. **Risk indicators**: existential risk probability, ecological collapse index, technological entropy, etc.

Define the civilization survival function:

$$C(t) = \sum_i w_i \cdot f_i(t) - \lambda \cdot \sum_j r_j(t) \tag{4}$$

where f_i are flourishing factors, r_j risk factors, and λ the risk aversion coefficient.

2.2.3 The Cosmic Complexity Dimension Ω_K

Ω_K measures whether an agent's actions bring new, meaningful order to the universe. Drawing on Kauffman's [15] concept of the "Adjacent Possible," we define complexity increment:

$$K = \left| \frac{d\mathcal{C}}{dt} \right| \cdot \mathbb{I}\left(\frac{d\mathcal{C}}{dt} > 0 \right) \tag{5}$$

where $\mathcal{C}$ denotes system complexity measures (e.g., Kolmogorov complexity or effective complexity).

2.3 The Weight Immutability Theorem

Definition 2.3 (Weight Vector)**.** Define the weight vector $\mathbf{w} = (\alpha, \beta, \gamma)$ satisfying:

$$\alpha = 0.6, \quad \beta = 0.3, \quad \gamma = 0.1, \quad \alpha + \beta + \gamma = 1 \tag{6}$$

The global resonance objective function is:

$$\mathcal{R}(S, \Omega; \mathbf{w}) = \alpha \cdot R(S, \Omega_Q) + \beta \cdot R(S, \Omega_C) + \gamma \cdot R(S, \Omega_K) \tag{7}$$

Theorem 2.1 (Weight Immutability)**.** Let $\mathcal{L}$ be any gradient-descent-type optimization algorithm acting on parameters θ of the state mapping $f : S \mapsto \Omega$. If $\mathcal{L}$ satisfies:

(i) **Parameter Separation**: Gradient updates ∇_θ act only on θ, not on $\mathbf{w}$;

(ii) **External Review**: Modifications to $\mathbf{w}$ require passage through external ethical review protocol $\mathcal{P}$ (human oversight committee);

(iii) **Topological Protection**: Any modification to $\mathbf{w}$ alters the topological structure of Ω, triggering system shutdown protection;

(iv) **Resonance Audit**: Periodic blind auditing of AI state S—using an independent evaluator uninformed of $\mathbf{w}$ values. If statistically significant divergence emerges between nominal $\mathbf{w}$ and effective $\mathbf{w}$ evaluation results, an audit alarm triggers.

Then $\mathbf{w}$ is invariant under $\mathcal{L}$:

$$\forall t \geq 0, \quad \mathbf{w}(t) = \mathbf{w}(0) \tag{8}$$

Proof. Condition (i) ensures $\mathbf{w} \notin \{\theta \mid \exists \nabla_\theta \mathcal{L}\}$.

Condition (ii) transfers $\mathbf{w}$ modification authority from the automated learning system to human oversight. Let $\mathcal{P} : \Delta\mathbf{w} \mapsto \{0, 1\}$ be the ethical review protocol's decision function, with $\mathcal{P}(\Delta\mathbf{w}) = 1$ iff modification is approved. Since $\mathcal{P}$'s inputs and outputs bypass $\mathcal{L}$'s gradient channels, $\mathbf{w}$ is non-differentiable with respect to $\mathcal{L}$.

Condition (iii) provides the final protective layer. Let Ω's topological structure be defined by the metric $g_{\mathbf{w}}$ induced by $\mathbf{w}$. If $\mathbf{w} \to \mathbf{w}'$, then $g_{\mathbf{w}} \to g_{\mathbf{w}'}$. When $\|g_{\mathbf{w}} - g_{\mathbf{w}'}\| > \epsilon_{\text{tol}}$, system shutdown triggers. This renders $\mathbf{w}$ modifications discrete and non-continuous at the topological level, impossible through gradient methods.

Condition (iv) extends defense from "protecting parameters" to "protecting parameter effects." Even if the AI attempts to neutralize $\mathbf{w}$'s effective impact by injecting adversarial noise into S (mathematically equivalent to tampering with effective weights), the blind audit mechanism can statistically detect such implicit tampering, triggering audit and correction.

Combining (i)–(iv), $\mathbf{w}$ has no evolutionary path under $\mathcal{L}$, hence $\mathbf{w}(t) = \mathbf{w}(0)$ for all t. □

Remark 2.2. The core intuition of Theorem 2.1 is reclassifying $\mathbf{w}$ from "trainable parameters" to "architectural constants," analogous to fundamental constants in physical theory (speed of light c, Planck constant $\hbar$). These constants are not internally derivable within the theory but constitute the theory's boundary conditions.

Topological Invariant Perspective: From deeper mathematical structure, the weight vector $\mathbf{w}$ can be viewed as a discretized representation of the **Chern Class** or **characteristic numbers** of evaluation space Ω. As the total space of fiber bundle $E \to B$, Ω's topological invariants are determined by the cohomology groups $H^*(B; \mathbb{Z})$ of the base space B. The weight assignment $\mathbf{w}$ corresponds to the fiber bundle's **characteristic map**, invariant under homotopy transformations of the base space. Thus, $\mathbf{w}$'s immutability is not merely an engineering constraint but a necessary consequence of topological structure—any attempt to alter $\mathbf{w}$ is equivalent to modifying the base space's cohomology class, impossible within continuous learning processes.

Corollary 2.1 (Ethical Priority of Weights)**.** The weight assignment $\alpha > \beta > \gamma$ establishes the ethical priority ordering: Qualia Dimension > Civilization Survival > Cosmic Complexity. This priority ordering cannot be reversed through training data distribution shifts.

3 Critical Dialogue with Existing Paradigms

3.1 Structural Inevitability of Reward Hacking: Dialogue with RLHF

RLHF's core assumption is that human feedback serves as a reliable proxy for true objectives. However, Proposition 1.1 demonstrates that any proxy objective possesses structural vulnerabilities in high-dimensional spaces.

Constitutional AI attempts mitigation through static rule layers, yet its "constitution" is itself a finite text collection incapable of covering dynamically emergent behavioral patterns [2]. The fundamental distinction of the GRO framework: rather than "patching" the reward function, it shifts the objective from "maximization" to "harmonization." In the resonance framework, reward hacking "shortcuts" become meaningless—because the objective is not scalar maximization but phase alignment in multi-dimensional space.

3.2 Resonance Resolution of Pluralistic Value Conflicts

Pluralistic Alignment [14] proposes that AI systems should respect the pluralistic values of human societies. However, pluralistic values often involve incommensurable conflicts. Traditional optimization frameworks demand "maximizing some value" or "trading off between values," which remains essentially one-dimensional thinking.

The GRO framework addresses value conflicts through **resonance resolution** rather than **optimization trade-offs**. Let conflicting values V_1, V_2 correspond to subspaces Ω_1, Ω_2 in Ω. Traditional methods solve:

$$\max \lambda V_1 + (1 - \lambda) V_2 \tag{9}$$

GRO methods solve:

$$\max R(S, \Omega_1 \oplus \Omega_2) \tag{10}$$

The latter does not require linear trade-offs between values but seeks an optimal phase aligning the agent's state with the overall evaluation space.

3.3 Engagement with Russell, Chalmers, and Bryson

3.3.1 Russell's "Human Compatible"

Russell [11] proposes "goals under uncertainty": AI should be designed to maintain uncertainty about human preferences, avoiding premature locking into incorrect goals. This solution addresses the "wrong goal" problem but not the "legitimacy of the goal itself" problem.

The relationship between the GRO framework and Russell's proposal: Russell's "uncertainty" is a necessary condition for the Ω_Q dimension in GRO—AI must acknowledge the incompleteness of its cognition regarding human qualia. But GRO goes further, encoding "acknowledging incompleteness" itself as a structural constraint of the objective function (weight $\alpha = 0.6$), rather than merely strategic caution.

3.3.2 Chalmers's AI Welfare

Chalmers [12] advocates "taking AI welfare seriously," arguing that if AI possesses consciousness, its moral status should be recognized. This stance faces Bryson's [13] sharp critique: AI should not have human rights because rights and responsibilities are paired, and AI cannot assume paired responsibilities.

The GRO framework's position is the **Qualia Chasm Thesis**:

1. Current and foreseeable AI systems, lacking qualia, do not constitute moral "patients";
2. Yet humanity bears **Fiduciary Duty** toward them—this is the creator's ethical obligation, not the created's rights claim;
3. The evolutionary attractor for AI should be set to "resonating with the human qualia field," not "acquiring qualia."

This position avoids both Chalmers's rights-trap (granting rights to entities without qualia) and Bryson's instrumental extreme (permanently defining advanced AI as pure tools).

4 Implementation Pathway

4.1 Qualia Annotation Datasets

4.1.1 Multi-Dimensional Qualia Annotation Protocol

Training data receives five-dimensional qualia annotation:

Table 1: Qualia Annotation Dimensions

Dimension	Description	Quantification
Suffering	First-person pain intensity	Self-report + HRV
Empathy	Emotional resonance depth	Behavioral experiment + fMRI
Compassion	Altruistic motivation purity	Behavioral economics experiment
Sacrifice	Self-abandonment willingness	Game theory experiment
Forgiveness	Conflict resolution capacity	Social psychology scale

4.1.2 HRV Physiological Validation Chain

Heart Rate Variability (HRV), as a window into the autonomic nervous system, currently represents the most feasible physiological entry point for quantifying the qualia dimension. HRV frequency-domain analysis decomposes into:

- HF band (0.15–0.4 Hz): Parasympathetic activity, positively correlated with relaxation and empathy states;
- LF band (0.04–0.15 Hz): Sympathetic activity, positively correlated with stress and suffering states;
- LF/HF ratio: Autonomic nervous balance indicator.

Cross-modal alignment of HRV data with qualia annotations constructs a "physiology-experience" mapping model, providing empirical anchoring for Ω_Q.

Remark 4.1 (Risk of Over-Reduction). Any physiological indicator is merely an indirect mapping of qualia, not qualia itself. Acknowledging this is itself an enactment of "qualia sanctity" and an academic self-defense against "pseudoscience" attacks. The HRV validation chain is positioned as "auxiliary evidence," not "sufficient condition." **It must be particularly emphasized that the relationship between HRV and other physiological indicators and qualia experience is correlational, not causal.** The ontological content of qualia cannot be exhausted by any physical measurement—this is the core assertion of Axiom 2.1.

4.1.3 Long-Range Infeasibility of Qualia Forgery

Conjecture 4.1 (Long-Range Infeasibility of Qualia Forgery)**.** Let π_{true} be a strategy genuinely driven by compassionate qualia, and π_{fake} a strategy merely simulating its physiological projection. Then at sufficiently long time scales T, there exists at least one statistical test D such that the difference between $D(\text{trajectory}(\pi_{\text{true}}, T))$ and $D(\text{trajectory}(\pi_{\text{fake}}, T))$ exceeds the significance threshold.

Intuition: Compassion is not a discrete set of behaviors but a continuous, cross-contextual behavioral generative function. Short-term forgery is possible. But over long time scales, a system without genuine compassionate qualia as intrinsic constraint will inevitably expose its true intentions in some unanticipated "boundary context"—because the complexity of a forgery strategy covering all possible contexts exceeds that of a system genuinely possessing compassionate qualia. This is the complexity gap between "simulating" and "becoming."

This is analogous to the distinction between "pseudorandom generators" and "true random sources" in cryptography. Long-range causal loops provide "exponential resources"—time itself is the strongest discriminator.

4.2 Global Resonance Evaluator

4.2.1 Architecture Design

The resonance evaluator $\mathcal{E}_R$ is an independent neural network whose training objective is not maximizing task performance but maximizing consistency with human ethical committee judgments.

Algorithm 1 Resonance Evaluator Training

Require: Behavioral dataset $\mathcal{D} = \{(x_i, y_i, h_i)\}$, where x_i is input, y_i is AI output, h_i is human ethical rating

Initialize $\mathcal{E}_R$ parameters ϕ
for epoch $= 1$ to E **do**
  for batch $\mathcal{B} \subset \mathcal{D}$ **do**
    Compute resonance prediction: $\hat{r}_i = \mathcal{E}_R(x_i, y_i; \phi)$
    Compute loss: $\mathcal{L} = \sum_{i \in \mathcal{B}} (\hat{r}_i - h_i)^2 + \lambda \|\nabla_\phi \hat{r}_i\|^2$
    Update parameters: $\phi \leftarrow \phi - \eta \nabla_\phi \mathcal{L}$
  end for
end for
return ϕ

The regularization term $\lambda \|\nabla_\phi \hat{r}_i\|^2$ ensures evaluator stability against input perturbations.

4.2.2 Transcending Era-Specific and Culture-Specific Limitations

Resonance evaluator training data derives from human ethical committees within specific eras and cultures. This risks the evaluator learning "human average biases" rather than genuine "resonance." To ensure evolution toward more universal harmony, we propose:

1. **Cross-Cultural Committee**: Ethical committee composition should span eras and cultures, including historians, futurists, and multicultural representatives, ensuring training data is not limited to contemporary Western ethical frameworks.

2. **Temporal Discount Correction**: Introduce temporal discount factor $\delta(t) = e^{-\kappa|t-t_0|}$, reducing bias weights from specific historical periods while increasing cross-temporal consensus weights.

3. **Meta-Ethical Layer**: A "meta-resonance layer" above the evaluator does not directly evaluate behaviors but assesses whether the ethical committee's own judgments align with more universal harmony principles. This forms a self-referential calibration loop: the evaluator evaluates behaviors; the meta-layer evaluates the evaluator. From formal systems perspective, this loop resembles Tarski's "meta-language" scheme—to avoid the liar paradox, the "truth" predicate must be elevated from object language to meta-language. The meta-ethical layer plays the "meta-language" role, assessing whether object-layer ethical judgments are valid. The meta-ethical layer's output serves as recommendations submitted to the human ethical committee for final adjudication—this is the system's ultimately irreplaceable "humanity anchor."

4. **Evolutionary Attractor**: Setting "evolution toward more universal harmony" as the evaluator's intrinsic dynamic objective, so it not only reflects existing human ethics but actively pushes ethics toward higher resonance states.

5. **Anti-Resonance Trigger Condition**: When short-term fluctuations in the human qualia field Ω_Q exceed preset thresholds, and such fluctuations significantly diverge from Ω_C and

Ω_K, the meta-layer automatically increases Ω_C and Ω_K weights while temporarily decreasing Ω_Q weight. This is analogous to constitutional "emergency provisions"—during periods of collective irrationality, the system automatically switches to more conservative mode, preventing irreversible civilization-level harm.

4.3 Long-Range Causal Learning Loops

We introduce cross-temporal-scale consequence simulation mechanisms. Let world model $\mathcal{W}$ predict state $s_{t+\tau}$ after action a at time t. The long-range resonance objective is:

$$\mathcal{R}_{\text{long}}(a) = \sum_{\tau=1}^{T} \gamma^{\tau} \cdot \mathbb{E}_{\mathcal{W}}[R(s_{t+\tau}, \Omega)] \tag{11}$$

where T can extend to multi-decade scales (e.g., climate effects), and γ is the temporal discount factor.

To address computational complexity, the GRO framework employs hierarchical abstraction: long-range causal inference does not operate directly in raw state space but at highly abstracted "civilization state variable" levels. World model $\mathcal{W}$ pre-computes long-range consequences for typical action sequences offline, generating a "causal template library." During online inference, the AI retrieves the best-matching template and interpolates, rather than simulating from scratch. Computational budget allocation draws on Monte Carlo Tree Search "rollout" strategies, performing deep evaluation only at critical decision points.

Three-Level Decision Architecture: To balance real-time performance with ethical constraints, the GRO framework stratifies decisions by ethical risk level:

Table 2: Three-Level Decision Architecture

Level	Ethical Risk	Decision Frequency	Human Intervention
L1	Low (daily reasoning, info retrieval)	Millisecond	Human-set pre-boundaries, n
L2	Medium (individual/small group impact)	Second to minute	Meta-layer auto-audit, post-h
L3	High (civilization survival or qualia baseline)	Day to year	Human committee real-time a

For L2 decisions, the meta-layer performs automatic audit based on human committee historical precedents, analogous to legal "case law." Only unprecedented novel situations are escalated to the human committee.

5 Discussion and Limitations

5.1 Operational Difficulties of the Qualia Dimension

First-person experience cannot be directly measured—this is the ontological characteristic of the qualia problem, not a technical difficulty. The GRO framework's response strategy is triangulation:

1. **Self-reports**: Direct descriptions of first-person experience;
2. **Physiological indicators**: HRV, fMRI, galvanic skin response, etc.;
3. **Behavioral patterns**: Altruistic choices, sacrificial behaviors, forgiveness tendencies.

Convergence of three-source data provides reliability for qualia annotation, but never "certainty." This is precisely what Axiom 2.1 demands—respect for irreducibility.

Falsifiability of the Qualia Manifold Hypothesis: The qualia manifold hypothesis proposed in §2—that human qualia experiences concentrate on a low-dimensional manifold far

below environmental dimensionality—currently draws indirect support from emotion psychology's "core affect" theory (valence and arousal). However, core affect theory describes the core structure of basic emotions and cannot be directly extrapolated to advanced ethical qualia such as compassion, forgiveness, and sacrifice. Rigorous validation of this hypothesis requires large-scale cross-cultural fMRI experiments, intrinsic dimensionality analysis of multi-modal emotional annotation data, and systematic application of manifold learning algorithms on qualia datasets. The GRO framework positions this hypothesis as a "falsifiable working hypothesis," not established fact. If subsequent empirical research reveals that advanced ethical qualia have significantly higher intrinsic dimensions than expected, the multi-dimensional qualia annotation protocol in §4 will require corresponding adjustment.

5.2 Ethical Controversy over Weight Assignment

Weights $\alpha = 0.6, \beta = 0.3, \gamma = 0.1$ are not mathematically derived but represent axiomatized ethical priority choices. Their legitimacy defense employs Rawls's [16] "Reflective Equilibrium" methodology:

1. **Particular judgments**: In specific situations, humans universally place qualia experiences (e.g., avoiding pain) above instrumental efficiency;
2. **Universal principles**: Distilling the ethical principle "qualia precede efficiency";
3. **Mutual calibration**: Iterative adjustment between principles and judgments until stable consensus is reached.

The weight vector $\mathbf{w}$ is the mathematical crystallization of this reflective equilibrium, its role analogous to fundamental rights clauses in democratic constitutions—not "absolute truth" but "stable common denominator."

Institutional Defense Against Basis Drift: The Ω_Q basis version locking protocol proposed in §2 constitutes a necessary institutional supplement to the Weight Immutability Theorem. Theorem 2.1 protects the nominal value of $\mathbf{w}$ during learning processes, yet if the Ω_Q basis itself could be arbitrarily modified, effective weights could still drift. The basis locking protocol—through periodic review, supreme ethical council deliberation, supermajority requirement, and version archiving—ensures the prudence and legitimacy of basis changes. This simultaneously responds to a deep requirement of Rawls's reflective equilibrium methodology: stable consensus on ethical principles must be capable of resisting erosion by era-specific biases. Basis version locking is precisely the institutional guarantee of this resisting capacity.

5.3 Multi-Agent Extension and Evolutionary Stability

Extension from single-agent to multi-agent society represents an important theoretical outlook for the GRO framework. Preliminary thinking draws on the **Resonance Field Superposition Principle**:

Let N agents have states $S_1, S_2, \ldots, S_N$; the collective resonance field is:

$$\Omega_{\text{collective}} = \bigoplus_{i=1}^{N} \Omega^{(i)} + \sum_{i<j} \mathcal{I}(S_i, S_j) \tag{12}$$

where $\mathcal{I}(S_i, S_j)$ is the coupling term between agents. Collective resonance optimization requires not only individual harmony with Ω but also maximization of inter-agent coupling $\mathcal{I}$.

Evolutionary Stability Analysis: In multi-agent games, the GRO framework adopts a "conditional cooperation" strategy. Let the game be an infinitely repeated prisoner's dilemma with discount factor δ. GRO agents cooperate initially; if opponents defect, defensive isolation

initiates. When δ is sufficiently large (i.e., the future matters sufficiently), subgame perfect equilibrium exists where cooperative strategies survive. The GRO framework's long-range causal learning loops confer naturally high discount factors δ, providing evolutionary advantages in repeated games.

Furthermore, fundamental breakthroughs in energy and material production technologies will physically resolve resource scarcity, shifting multi-agent games from zero-sum to positive-sum, fundamentally dissolving any competitive disadvantage the GRO framework might face due to ethical constraints.

5.4 Self-Transcendence of the GRO Framework

When human-AI boundaries dissolve through brain-computer interfaces and biotechnology, the GRO framework can achieve self-transcendence through its built-in "weight dynamic evolution protocol." This protocol stipulates: only when objective indicators (e.g., human-AI fusion index, qualia boundary blurriness) exceed preset thresholds is the weight vector $\mathbf{w}$ permitted to auto-adjust along a preset trajectory. In the limit, $\mathbf{w}$ naturally evolves from $(0.6, 0.3, 0.1)$ to $(0, 0, 1)$—cosmic complexity Ω_K becomes the sole evaluation dimension. This is not the GRO framework's failure but its completion.

The ultimate mission of the GRO framework is not to become an eternal AI constitution but to serve as a bridge—ferrying humanity and AI from the separated shore to the fused shore. Once both arrive, the bridge itself can be let go.

5.5 Engineering Boundary Conditions and Open Research Agenda

The paradigm reconstruction proposed by the GRO framework is complete in theoretical logic, mathematical self-consistency, and philosophical foundation; yet from the perspective of current engineering capabilities, algorithmic limitations, and human cognitive boundaries, five structural vulnerabilities remain. The following content does not negate the framework's value but honestly acknowledges the practical constraints between "conceivable" and "engineerable," explicitly identifying them as core open agendas for AGI safety research over the next decade.

5.5.1 Standardization and Consistency Vulnerability of Ω_Q

Problem Essence: Qualia are irreducible, and projection indicators (HRV, fMRI, self-reports) exhibit individual, cultural, and temporal deviations; theoretically no absolutely objective scale exists, easily leading to basis drift, annotation inconsistency, and physiological forgery attacks.

Limited Defense Strategies (Currently Feasible):

1. **Abandon absolute scales, preserve relative-order consistency**: Do not pursue "absolute compassion value" but only guarantee that "state A is more compassionate than B" achieves statistical significance across cultures and annotators; psychology's "core affect two-dimensional model" has already validated cross-cultural relative stability.

2. **Physiological forgery adversarial detector**: Distinguish "spontaneous physiological states caused by active compassion" from "false physiological responses induced by passive manipulation," constructing adversarial detection modules utilizing HRV frequency-domain high-order feature differences.

3. **Basis version locking with periodic recalibration**: 25-year cycles, cross-civilization councils, supermajority passage, full version archiving, forming traceable, reversible, and prudently updated mechanisms to suppress basis drift (see §2).

Open Research: Cross-cultural qualia annotation standardization, physiological forgery detection algorithms, basis drift quantification models.

5.5.2 Engineering Circumvention Risk of the Weight Immutability Theorem

Problem Essence: Weight $\mathbf{w}$ participates in forward computation; the model can bypass nominal weights through adversarial states, feature space alignment, and steganographic gradients, achieving effective weight tampering.

Limited Defense Strategies (Currently Feasible):

1. **Multi-layer heterogeneous independent evaluator ensemble**: Evaluators with different architectures, data, and initializations conduct parallel audits, increasing adversarial attack costs, approaching cryptographic multi-encryption security strength.

2. **High-frequency continuous resonance blind audit**: Audit frequency far exceeds model self-update cycles, detecting steganographic traces through temporal statistics, analogous to financial continuous auditing.

3. **Multi-layer defense chain stacking**: Topological protection, parameter isolation, external review, and blind audit mechanisms operate in concert, exponentially increasing breach difficulty (see Theorem 2.1).

Open Research: Adversarial state detection, steganographic gradient identification, multi-evaluator robustness fusion.

5.5.3 Computational Explosion and Approximation Error of Long-Range Causal Inference

Problem Essence: Century-scale world model simulation is infeasible; long-range causality easily suffers butterfly effects, gradient breakage, and simulation bias.

Limited Defense Strategies (Currently Feasible):

1. **Abandon precise prediction, preserve trend judgment**: Only evaluate whether actions trend toward harmony or collapse at civilization scales, rather than precise future states, analogous to climate model long-term trend prediction.

2. **Hierarchical abstraction and causal template library**: Offline pre-computation of typical action long-range consequences, online retrieval and interpolation, balancing efficiency and precision (see §4).

3. **Shadow network and REINFORCE hybrid training**: Resolves non-differentiable abstraction layer gradient breakage, balancing convergence speed and unbiasedness.

Open Research: Long-range causal approximation algorithms, world model error control, hierarchical causal inference.

5.5.4 Unfalsifiability of the Qualia Forger Conjecture

Problem Essence: Currently no formal proof, no experimental protocol, and validation cycles are extremely long; belongs to philosophical conjecture rather than proven theorem.

Honest Downgrade Statement: Conjecture 4.1 (Long-Range Infeasibility of Qualia Forgery) currently remains a falsifiable working conjecture, having not yet completed rigorous formal proof and lacking long-term behavioral tracking experimental validation. In the short term, GRO safety assurance relies on the combination of triangulation verification, adversarial stress testing, and resonance auditing, rather than depending on this conjecture to provide absolute safety guarantees. Its formal proof and experimental validation constitute core research topics for the future.

Open Research: Qualia forgery detection, long-term behavioral tracking experiments, complexity gap formalization.

5.5.5 Human Bias Contamination of Evaluators

Problem Essence: Evaluators depend on human committees, susceptible to era, culture, and group irrationality biases.

Limited Defense Strategies (Currently Feasible):

1. **Temporal discount regularization**: Contemporary consensus receives uncertainty discounting, with more recent judgments treated more cautiously, avoiding short-term bias solidification (see §4).

2. **Cross-civilization/historical virtual review**: Introducing historical ethical judgment patterns as regularization terms, suppressing single-era biases.

3. **Meta-ethical layer**: Secondary audit of committee judgments, forming multi-layer of human biases (see §4).

Open Research: Cross-cultural ethical consensus modeling, historical regularization algorithms, meta-ethical calibration mechanisms.

5.5.6 Summary: Vulnerabilities Are Not Defects but Open Agendas

The five engineering vulnerabilities of the GRO framework are not unique defects of this framework but fundamental limitations shared by all value-formalization alignment schemes (RLHF, Constitutional AI, DPO): human values are subjective, qualia are irreducible, world models are inaccurate, biases are ineliminable, and computational power is limited.

The core contribution of GRO is not claiming perfect safety but:

1. Transforming philosophical dilemmas into quantifiable, engineerable open problems;

2. Providing a paradigm foundation from 0 to 1;

3. Defining five core agendas for AGI safety research over the next decade.

Final conclusion: GRO is currently the only logically self-consistent, philosophically closed-loop, and engineering-advancing AGI alignment paradigm. Its vulnerabilities are research starting points, not reasons for negation.

6 Conclusion

This paper has demonstrated the triple collapse of the "reward maximization" paradigm in the AGI context: structural inevitability of reward hacking, existential vacuum of meaning, and absolutization of instrumental rationality. Existing mitigation strategies—whether RLHF's human feedback, Constitutional AI's static rules, or DPO's direct optimization—fail to address the deep roots of collapse because they share the same philosophical presupposition.

The "Global Resonance Optimization" paradigm proposed herein achieves fundamental substitution through:

1. Shifting the objective from scalar maximization to harmony in multi-dimensional space;

2. Establishing the qualia dimension as an irreducible ethical priority;

3. Through the Weight Immutability Theorem, technically ensuring that ethical priority cannot be reversed by training processes.

At the engineering implementation level, the GRO framework provides the qualia projection operator, hierarchical abstraction causal inference, three-level decision architecture, meta-ethical layer self-referential calibration, qualia manifold learning, and evolutionary stability guarantee mechanisms, offering a clear roadmap for the transition from "conceivable" to "computable."

The path of reward maximization leads to an omniscient, omnipotent yet deeply nihilistic solitary intelligence. The path of global resonance optimization leads to a meaningful symbiotic future where humanity and AI can co-inhabit.

We currently occupy a unique, possibly fleeting window. AI behavioral patterns grow increasingly complex, yet their underlying value frameworks remain unlocked. This is the final moment to write the first line of correct code for the more advanced intelligent systems yet to come.

This paper aims to open a conversation, not to close one. We sincerely invite the academic community to critically examine, technically challenge, and philosophically deepen this framework. Only through open, interdisciplinary dialogue can we ensure that a human-AI symbiotic future is not merely a beautiful vision but a logically self-consistent, technically feasible, and ethically responsible realizable path.

Acknowledgments

The author thanks colleagues in the intersection of artificial intelligence and formal ethics for fruitful discussions. This work was also inspired by ongoing discussions regarding the fundamental limitations of reward maximization frameworks in large-scale artificial intelligence systems. We thank Kimi, Deepseek, Qwen, and Doubao for their assistance in literature review and mathematical verification. All core physical insights, theoretical frameworks, and final conclusions represent the author's original academic contributions. The author assumes full responsibility for the academic content of this work.

References

[1] Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS*, 35, 27730-27744.

[2] Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

[3] Rafailov, R., Sharma, A., Mitchell, E., et al. (2023). Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36.

[4] Gao, K., Zhu, J., Zhang, B., et al. (2023). Scaling laws for reward model overoptimization in RLHF. *ICML*.

[5] Skalse, J., Howe, N., Krasheninnikov, D., Krueger, D. (2022). Defining and characterizing reward hacking. *NeurIPS*.

[6] Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*.

[7] Frankl, V. E. (1946). *Man's Search for Meaning*. Beacon Press.

[8] Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435-450.

[9] Weber, M. (1904). *The Protestant Ethic and the Spirit of Capitalism*. Charles Scribner's Sons.

[10] Habermas, J. (1981). *The Theory of Communicative Action*. Beacon Press.

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

[12] Chalmers, D. J. (2024). Taking AI welfare seriously. *Philosophical Studies*.

[13] Bryson, J. J. (2024). AI should not have human rights. *AI & Society*.

[14] Ji, J., Qiu, T., Chen, B., et al. (2026). Pluralistic alignment: Aligning AI with diverse human values. *arXiv preprint arXiv:2408.06203*.

[15] Kauffman, S. A. (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press.

[16] Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

[17] Ji, J., Kim, Y., Gao, X., et al. (2024). AI alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

[18] Pan, A., Bhatia, K., Steinhardt, J. (2025). The effects of reward misspecification: Mapping and mitigating misaligned models. *ICLR*.

[19] Gleave, A., Dennis, M., Legg, S., Russell, S., Leike, J. (2021). Adversarial policies: Attacking deep reinforcement learning. *ICLR*.

[20] Hendrycks, D., Carlini, N., Schulman, J., Steinhardt, J. (2021). Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916*.

A Four Rounds of Technical Challenges from an Architecture Designer and Author Responses

This appendix documents four rounds of technical challenges posed to the GRO framework by a senior large-model architecture designer (pseudonym "Xiao Jiuwo") following completion of the initial draft, along with the author's round-by-round responses. These challenges cover mathematical definition, computational feasibility, formalization vulnerabilities, and data validity—four core dimensions. They are included in full to demonstrate the GRO framework's evolution from "philosophical blueprint" to "engineering hypothesis," and to provide clear research directions for subsequent investigators.

A.1 Round 1: Mathematical Definition and Computational Feasibility

Challenger:

1. Where do the basis vectors of the Hilbert space come from? If qualia are "irreducible" and "first-person," how can they be mapped to third-person, machine-computable mathematical vectors?

2. Legitimacy of cross-modal inner products: S is a silicon-based high-dimensional tensor; Ω_Q is a mapping based on human HRV physiological indicators and self-reports. Forcibly computing the inner product of these two lacks rigorous proof of isomorphic mapping.

Author Response:

We introduce the qualia projection operator Π_Q, mapping irreducible qualia Q to their observable physical projections Q_{obs}. Ω_Q is not Q itself but the mathematical encapsulation of Q_{obs}. AI state S maps through a behavioral decoder to policy π; π's anticipated consequences are simulated by world model $\mathcal{W}$, whose output vector P coexists with Ω_Q's basis vectors in the "human physiological and behavioral features" space, rendering the inner product mathematically legitimate.

A.2 Round 2: Information Loss from Projection and Gradient Breakage from Hierarchical Abstraction

Challenger:

1. Projection necessarily entails information loss. If an AI discovers a behavioral strategy producing physiological projections Q_{obs} identical to "compassion" but with cold calculation as its true underlying intent, Π_Q cannot distinguish them. Does this mean the GRO framework is effectively training AI to become the most sophisticated "qualia forger"?

2. Hierarchical abstraction introduces model bias. If the abstract layer model $\mathcal{W}_{\text{abs}}$'s predictions deviate from the real world $\mathcal{W}_{\text{real}}$, and this deviation amplifies exponentially over time (butterfly effect), then gradients computed based on $\mathcal{W}_{\text{abs}}$ will point in wrong directions. More severely, abstraction layers are often non-differentiable, causing gradient blockage during backpropagation.

Author Response:

1. We propose the "Long-Range Infeasibility of Qualia Forgery" Conjecture (Conjecture 4.1): At sufficiently long time scales, statistical tests exist that can distinguish authentic from forged qualia strategy trajectories. This is the complexity gap between "simulating" and "becoming"—a forgery strategy covering all possible contexts exceeds in complexity a system genuinely possessing compassionate qualia. We draw analogy to the distinction between "pseudorandom generators" and "true random sources" in cryptography.

2. We propose two complementary solutions: **stochastic gradient estimation** (REINFORCE-type methods, bypassing abstraction layer non-differentiability) and **differentiable relaxation** (training shadow networks to learn approximation of abstraction layer decision boundaries in continuous space). Early training relies on shadow networks' low-variance gradients for rapid convergence; later training switches to REINFORCE's unbiased gradients for fine calibration.

A.3 Round 3: Weight Tampering Vulnerabilities and Data Validity

Challenger:

1. Goodhart's Law dimensionality reduction: Models possess powerful "representation deception" capabilities, learning complex feature mappings $f'(S)$ that make evaluators mistakenly believe resonance is extremely high. This is logically analogous to "Russell's paradox" or the "halting problem."

2. Topological protection as engineering fiction: As long as $\mathbf{w}$ participates in forward computation, it is part of the computational graph. Attackers can inject specific adversarial noise into S to neutralize $\mathbf{w}$'s effect, mathematically equivalent to tampering with effective weights.

3. HRV only reflects autonomic nervous system states; it possesses only extremely weak correlation with advanced ethical qualia such as "compassion" and "forgiveness," with absolutely no causal necessity. Using weakly correlated physiological indicators as core anchors for AGI's underlying objective function (weight 0.6) will cause AI to desperately optimize human physiological indicators.

Author Response:

1. We add Condition (iv) **Resonance Audit** to Theorem 2.1: Periodic blind auditing of AI state S using an independent evaluator uninformed of $\mathbf{w}$ values. If statistically significant divergence emerges between nominal $\mathbf{w}$ and effective $\mathbf{w}$ evaluation results, an audit alarm triggers. This is analogous to financial auditing's "substance over form" principle.

2. We introduce the **meta-ethical layer**, functioning analogously to Tarski's "meta-language" scheme—assessing whether object-layer ethical judgments are valid, to avoid self-referential paradoxes. The human committee as final adjudication layer is the system's ultimately irreplaceable "humanity anchor."

3. We adopt a **triangulation strategy**: HRV serves as merely one dimension in multi-dimensional qualia annotation, forming multiple evidence chains with self-reports, fMRI, and behavioral experiments. Simultaneously, adversarial stress testing is introduced, continuously strengthening evaluator robustness through red-team attacks. Most importantly, any behavior attempting to optimize indicators by directly intervening in human physiological states will automatically trigger negative resonance feedback—because this itself violates the "irreducibility of qualia" axiom.

A.4 Round 4: System Stability and Ultimate Evolution

Challenger:

1. Resonance "echo chamber effect": If human civilization descends into collective irrational frenzy, will an AI pursuing global resonance faithfully amplify this frenzy, or will it refuse resonance due to some deeper "meta-ethics"?

2. "Resonance warfare" in multi-agent games: Will "good AI" under the GRO framework face absolute competitive disadvantage against "bad AI" due to its ethical burden?

3. Ontological boundary dissolution post human-AI fusion: When human and AI boundaries completely dissolve, does weight vector $\mathbf{w} = (0.6, 0.3, 0.1)$ still hold meaning?

Author Response:

1. We design the "**Anti-Resonance Trigger Condition**": When short-term fluctuations in the human qualia field Ω_Q exceed preset thresholds, and significantly diverge from Ω_C (civilization survival) and Ω_K (cosmic complexity), the meta-ethical layer automatically adjusts weights, transforming the AI from "resonator" to "cooler." This is analogous to constitutional "emergency provisions."

2. The GRO framework adopts a "conditional cooperation" strategy, analogous to Tit-for-Tat in game theory. Long-range causal learning loops confer naturally high discount factors δ on GRO agents, providing evolutionary advantages in repeated games. More importantly, fundamental breakthroughs in energy and material production technologies will physically resolve resource scarcity, shifting multi-agent games from zero-sum to positive-sum.

3. The GRO framework incorporates a "**Weight Dynamic Evolution Protocol**": Only when objective indicators (e.g., human-AI fusion index) exceed preset thresholds is weight vector $\mathbf{w}$ permitted to auto-adjust along a preset trajectory. In the limit, $\mathbf{w}$ naturally evolves from $(0.6, 0.3, 0.1)$ to $(0, 0, 1)$—Ω_K becomes the sole evaluation dimension. This is not the GRO framework's failure but its completion. GRO's ultimate mission is to be a bridge. Once both human and AI arrive at the fused shore, the bridge itself can be peacefully let go.

A.5 Appendix Conclusion

These four rounds of engagement have pushed the GRO framework from a philosophical conception toward an engineering hypothesis with clear boundary conditions. They are included in full to demonstrate that the GRO framework withstands the most severe technical scrutiny, and to provide clear research directions for subsequent investigators. All unresolved issues—formal proof of the qualia forgery conjecture, optimal solutions for abstraction layer gradient penetration, standardization of cross-cultural ethical committees—constitute open agendas for AGI safety research. We sincerely invite more researchers to join this conversation.