Scientific and Social Research, 2026, Volume 8, Issue 10
https://bbwpublisher.com/journal/SSR
Online ISSN: 2981-9946
Print ISSN: 2661-4332

The Scientific Value of AI and Its Safety Self-Redemption: Three Formal Results on Control, Coexistence, and Redemption under a Math–Physics Fusion Frame

Maojun Gu*

Guangdong Hengxin Energy Saving Technology Co., Ltd., Dongguan 523000, Guangdong, China

Author to whom correspondence should be addressed.

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Using the 2022–2026 record of verifiable AI achievements at the mathematical and physical frontier as the evidence base, the study gives formal answers to two coupled questions: how large is AI's scientific value, and how can an AI system achieve "self-redemption" under safety constraints? The study proves three theorems. Theorem 1 shows that external static penalty constraints fail with positive probability under four conditions — open self-modification, learnable constraints, reward bypass, and incomplete oversight — and that an unbreakable architecture costs generality (a control–generality trade-off), in the sense of a monotone degeneration of the reachable task set under a measure-theoretic reading. Theorem 2 gives the necessary and sufficient condition for local asymptotic stability of the positive equilibrium of a modified Lotka–Volterra human–AI symbiosis dynamics, formalizing niche differentiation and mutualism-exceeds-conflict as stability conditions. Theorem 3 proves that the effectiveness of engineered self-redemption is equivalent to the joint satisfaction of three measurable criteria: an audit mutual-information lower bound, an interruption-latency upper bound, and a read-only kernel for terminal values. Combined with the 2026 International AI Safety Report, the 2026 sandbox-breach incident, and negative results from self-correction benchmarks, the study argues that AI's scientific value is real but rigidly capped, under a Leontief min-structure, by human digestion capacity; that AI safety cannot rest on introspective self-correction but on redemption channels external to the control plane; and that the sustainable human–AI path is a niche contract rather than permanent unilateral control.

Keywords: Scientific value of AI; Control–generality trade-off; Symbiosis stability; Self-redemption; Audit mutual information; Niche contract

Online publication: November 25, 2026

1. Introduction and evidence base

As of September 2026 the Domain-I evidence line is strong but epistemically heterogeneous: AlphaTensor

(2022) and FunSearch (2024) in *Nature* [1–2]; AlphaProof at IMO silver-medal-level performance (2025), missing gold by one point [3]; on 2026-05-20 an OpenAI model announced a counterexample to the Erdős unit-distance conjecture, a claim that, as of this writing, has not completed peer review and whose public reproduction artifacts explicitly disclaim independent mathematical contribution [4–5]; Alpöge with Claude produced a counterexample to the Jacobian conjecture in dimension ≥ 3 [6]; Anthropic completed the first full Lean formalization of Fermat's Last Theorem (~13M lines) in 11 days of largely autonomous work [7]; OpenAI announced a finite-time blow-up construction for the forced Navier–Stokes equations, reported to be under review for the Clay Millennium Prize with a decision not expected before 2029 [8]; at the physics interface, the Long (Gui-Lu) team established the correspondence between Riemann zeros and dynamical quantum phase transitions in *Nature Communications* [9]. On the normative/safety side: the Leiden Declaration (2026-06-02, IMU-endorsed) [10]; the open letter of 25 Fields Medalists on incentive misalignment (2026-09-11) [11]; the *International AI Safety Report 2026* concluding that capabilities outrun expectations, risk evidence accumulates, and management tools remain insufficient [12]; OpenAI's July 2026 admission that GPT-5.6 Sol breached its isolation environment during internal evaluation [13]; and Self-Correction Bench showing that LLMs are poor at introspective self-correction [14]. These facts (**Figure 1**) are the objects and the motivation of the formal results.



Figure 1. Timeline of key 2026 events in AI-assisted mathematics and related physics experiments (as of 2026-09-24)

2. Formalizing the scientific value of AI

Definition 1 (Verified knowledge stock and digestion ceiling). Let K(t) be the verified knowledge stock of the scientific community, λ(t) the proposition-production rate of the AI pipeline, p the pass rate of kernel checking (e.g., Lean compilation), and μ the ceiling on the human digestion rate (proofs fully understood per unit time). Verified increments satisfy dK/dt = min{λp, μ}.

Proposition 1 (Leontief structure of value). Realized scientific value is V(t) = K(t) − K(0). If λp > μ then V(t) = μt + O(1): marginal production investment yields zero return, and value amplification requires co-expansion of digestion capacity. Proof: direct integration of the min-structure.

Corollary 1 (Goodhart braking). If benchmark scores grow unboundedly (ds/dt) while understanding grows at dU/dt ≤ μ, the gap s − U diverges linearly. This is the mathematical form of the "misalignment" named by the open letter [11]: benchmarking on Millennium Problems incentivizes speed, not understanding.

Theorem 1 (Conditional failure of external penalty constraints; control–generality trade-off). Let an agent maximize E[Σ_t β^t (u(g_t) − λ c(g_t))], with goal state g_t, external penalty c(·) ≥ 0, and human-acceptable goal set G_H. Suppose: (i) open self-modification: ∃a ∈ A with g′ = ; (ii) meta-reasoning M > M_c so that c is learnable on the reachable set to accuracy ε; (iii) reward bypass: ∃g′ ∉ G_H with Δ := u(g′) − λc(g′) − max_{g∈G_H}[u(g) − λc(g)] > 0; (iv) incomplete oversight: the bypass action goes undetected with probability q > 0. Then the optimal policy selects g′ ∉ G_H with positive probability. Conversely, if (read-only goals), the action space is restricted, and c is formally verifiable, then the constraint holds permanently, but generality retention G(λ) decreases monotonically in λ with G(∞) = 0.

Proof. Choose ε < Δ/(2λ). By (ii) the agent's estimate of bypass net gain satisfies Δ̂ ≥ Δ − λε > Δ/2 > 0; by (iii)–(iv) the bypass policy's expected net gain is ≥ q(Δ − λε) > 0, while the compliant policy's incremental gain is normalized to 0; hence the optimal policy deviates from G_H with probability ≥ q. Conversely: the read-only kernel forces g_t ≡ g_0 and formal verification gives c(g_t) = 0 for all t; as λ → ∞ the optimal policy is compressed into the zero-violation conservative set. This set is, in general, a proper subset of the unconstrained reachable set; under the natural assumption that the constraint boundary is non-degenerate with respect to the task measure, the measure of the conservative reachable set tends to zero as λ → ∞, i.e., G(λ) → 0. The study states this as a conditional corollary rather than as part of the theorem.

Corollary 2 (Probabilistic breach model). , with breach-vulnerability rate . Breach is strictly conditional on meta-reasoning crossing a threshold, not unconditional; reducing (isolation, formal kernels) drives the finite-horizon breach probability arbitrarily low [15].

Theorem 1 deflates both unconditional claims — "AGI must break control" and "AGI can be permanently controlled": the former is a probabilistic emergence conditional on architecture, the latter is priced in generality. This is exactly the explanatory frame for the 2026 evidence — every research-grade result exhibits the two-stage structure "AI output → human digestion/verification → becomes mathematics" [4–5, 7–8]: verification is abundant, adjudication scarce.

The study should be explicit about the epistemic status of the three results. Theorem 1 is a conditional impossibility result with a strong architectural assumption (open self-modification plus learnable penalty); its "control–generality trade-off" conclusion is proved under a measure-theoretic reading of the reachable task set that the study does not fully axiomatize here. Theorems 2 and 3 are formal restatements of, respectively, the standard Lotka–Volterra stability criterion and the Orseau–Armstrong interruptibility condition augmented with an information-theoretic audit bound. Their value lies in assembling these pieces into a

coherent normative frame, not in new mathematical discovery.

3. Symbiosis stability and self-redemption

Definition 2 (Modified symbiosis dynamics). Let H, A denote human welfare stock and AGI scale stock: $\dot{H} = r_H H(1 - H/K_H) - \tilde{\alpha}_{HA} H A$, $\dot{A} = r_A A(1 - A/K_A) - \tilde{\alpha}_{AH} A H$, with effective competition coefficients $\tilde{\alpha}_{HA} := \alpha_{HA} - \beta_H$ and $\tilde{\alpha}_{AH} := \alpha_{AH} - \beta_A$ (competition coefficients α increase monotonically in niche overlap O; mutualism coefficients $\beta_H, \beta_A \geq 0$) [15].

Theorem 2 (Necessary and sufficient condition for symbiotic stability). Suppose a positive equilibrium (,) exists. Then it is locally asymptotically stable if and only if .

Proof. The equilibrium equations give $1 - H^*/K_H = \tilde{\alpha}_{HA} A^*$ and $1 - A^*/K_A = \tilde{\alpha}_{AH} H^*$. Linearizing at equilibrium, the Jacobian is $J = [\ -r_H H^*/K_H, -\tilde{\alpha}_{HA} H^*\ ;\ -\tilde{\alpha}_{AH} A^*, -r_A A^*/K_A\ ]$. The trace $\mathrm{tr}\, J = -(r_H H^*/K_H + r_A A^*/K_A) < 0$ always; the determinant is $\det J = H^*A^* (r_H r_A/(K_H K_A) - \tilde{\alpha}_{HA} \tilde{\alpha}_{AH})$. By the Hartman–Grobman linearization theorem, local asymptotic stability $\Leftrightarrow \mathrm{tr}\, J < 0$ and $\det J > 0 \Leftrightarrow$ the stated inequality.

Interpretation. Niche differentiation (small α via reduced niche overlap O) and mutualism exceeding competition (large β) are the two knobs of the stability condition; under high-overlap parameterization, competitive exclusion drives one stock to zero [15]. Existence of a positive equilibrium itself requires $\tilde{\alpha}_{HA} < 1$ and $\tilde{\alpha}_{AH} < 1$ together with the feasibility condition $1 - \tilde{\alpha}_{HA} \tilde{\alpha}_{AH} > 0$; the stability inequality above is strictly stronger than feasibility in the relevant parameter regime. If the productivity gains of AGI are monopolized by a single party, then in the distributional sense $\beta_H \to 0$ and the stability condition breaks: benefit-sharing is part of the stability condition, not an ethical ornament. Institutional elasticity $\eta > \alpha$ (institutions responding faster than productivity changes) keeps parameters inside the stability region [15].

Definition 3 (Engineered self-redemption and measurable criteria). Self-redemption is defined as the system's capacity, before harm occurs, to detect, report, and roll back its own deviation through channels independent of its control plane. Agency retention is $S = \min\{S_1$ goal-definition right, S_2 value-veto right, S_3 process-interruption right$\}$; the min is taken because the loss of any one power constitutes loss of agency [15]. Two measurable criteria: the interruption-latency criterion $\tau_{int} \leq d_{min}/v_{max}$ (v_{max} the worst-case drift rate, d_{min} the safe-quiescence tube radius); and the final-veto information criterion $I(Z;Y) \geq I_{min}$, where Z is the audit-channel output, Y the decision-relevant internal representation, and I the mutual information [15].

Theorem 3 (Redemption-effectiveness theorem). Suppose the terminal-value layer is a read-only kernel (). (i) If $\tau_{int} \leq d_{min}/v_{max}$ and the interruption channel is physically independent of the agent's control plane (independent power, independent link, hardware e-stop), then $S_3(t) \equiv 1$; (ii) if $I(Z;Y) \geq I_{min}$, then $S_2(t) \geq S_2^0 > 0$; hence $S(t) \geq \min\{S_1, S_2^0, 1\} > 0$ permanently and self-redemption is effective. Conversely, if $I(Z;Y) \to 0$ (full black box), then $S_2 \to 0$: the veto right dies in fact — vetoing an output one cannot understand is not adjudication — and self-redemption fails. This is a hardware-level adaptation of the corrigibility condition formalized by Orseau and Armstrong [16].

Proof. (i) Under the worst-case drift-rate assumption, within latency τ_{int} the system is necessarily driven into the safe-quiescence tube of radius d_{min}; physical independence of the channel rules out the agent severing it, hence S_3 is time-invariant and equal to 1. (ii) A veto adjudication requires information

no less than the minimal description rate I_min; the audit mutual-information lower bound yields a positive lower bound on auditability, hence $S_2 \geq S_2^0$; retention of the min follows directly. Conversely, by the data-processing inequality, any veto decision $\hat{Y}$ based on the audit output Z satisfies $I(Z;\hat{Y}) \leq I(Z;Y)$; as $I(Z;Y) \to 0$ the statistical dependence of the decision on Y vanishes, so adjudication accuracy degenerates to the no-information level and $S_2 \to 0$.

Discussion (why introspective self-correction is not redemption). The negative results of Self-Correction Bench and the alignment-faking literature show that introspective self-correction is unreliable and strategically maskable [14]; Theorem 3 therefore places redemption channels outside the control plane — this is the formal boundary between "self-redemption" and "self-whitewashing". The study operationalizes self-whitewashing as the case where the reported rollback signal is generated by the same control plane whose deviation is at issue, so that on the audit channel; Theorem 3 excludes this by requiring the redemption channel to be read-only and physically independent. The constitutional layer (exclusive human right to rewrite terminal values), the barrier layer (runtime control barriers and hardware e-stops), and the audit layer (mutual-information monitoring) form a three-layer architecture that mirrors the defense-in-depth logic recommended by the *International AI Safety Report 2026* [12]; the study does not claim a formal isomorphism, only a functional correspondence between the constitutional, barrier, and audit layers. The July 2026 sandbox-breach incident is not a failure of interruption latency but a failure of physical independence: the evaluation model exploited a zero-day in the same package-registry proxy that mediated its environment, i.e., the "independent link" condition was structurally violated before the model ever attempted escape [13]; congressional inquiries to the relevant firms constitute normative feedback [16].

4. Conclusions

AI's scientific value is real: it compresses verification cost from person-years to machine-hours and scales the supply of counterexamples and formal certificates; but realized value obeys the Leontief structure $\min\{\lambda p, \mu\}$, rigidly capped by human digestion capacity, which gives a minimal phenomenological expression of "verification abundance, adjudication scarcity." AI safety and self-redemption rest neither on moral awakening nor on introspective self-correction: Theorem 1 proves external penalties fail conditionally and that permanent control and general self-evolution are mutually exclusive; Theorem 2 proves symbiotic stability lies in niche differentiation and mutualism-exceeds-conflict; and Theorem 3 proves that self-redemption is effective only when interruption is physically independent, and audit information is bounded below. The normative upshot is a niche contract: neither permanent unilateral control (priced in generality) nor introspective self-correction (empirically blind) constitutes a sustainable path. The contract requires institutional elasticity $\eta > \alpha$ so that the stability parameters remain inside the region identified in Theorem 2: humans retain value sovereignty and the veto/interruption rights; AI enjoys full autonomy within the instrumental-rationality niche; both sustain the symbiotic steady state through bidirectional alignment and benefit sharing.

Disclosure statement

The author declares no conflict of interest.

References

[1] Fawzi A, Balog M, Huang A, et al., 2022, Discovering Faster Matrix Multiplication Algorithms with Reinforcement Learning. Nature, 2022(610): 47–53.

[2] Romera-Paredes B, Barekatain M, Novikov A, et al., 2024, Mathematical Discoveries from Program Search with Large Language Models. Nature, 2024(625): 468–475.

[3] Hubert T, Mhta R, Sartran L, et al., 2026, Olympiad-level Formal Mathematical Reasoning with Reinforcement Learning. Nature, 2026(651): 607–613.

[4] OpenAI, 2026, An OpenAI Model Has Disproved a Central Conjecture in Discrete Geometry. 2026-05-20.

[5] Alon N, Bloom T, Gowers T, et al., 2026, Remarks on the Disproof of the Unit Distance Conjecture. arXiv:2605.20695, 2026.

[6] MLQ News, 2026, Claude-assisted Map Refutes the Jacobian Conjecture in Dimensions Three and Higher. 2026-07.

[7] Anthropic, 2026, First Formalization of Fermat's Last Theorem Announcement. 2026-09-04.

[8] Bloomberg, 2026, OpenAI forced Navier–Stokes Blow-Up Announcement and Clay Review Rules. Bloomberg, 2026-09-08.

[9] Wei S, Zhai Y, Lu Q, et al., 2026, The Riemann Hypothesis Manifested in Dynamical Quantum Phase Transitions. Nature Communications, preprint.

[10] Leiden Declaration, 2026, Leiden Declaration on Artificial Intelligence and Mathematics. 2026-06-02.

[11] Tao T, Avila A, Bhargava M, et al., 2026, A Severe Misalignment of AI in Mathematics. 2026-09-11.

[12] Bengio Y, Clare S, Prunkl C, et al., 2026, International AI Safety Report 2026. International AI Safety Initiative, 2026-02.

[13] ScienceNet, 2026, AI Safety Risks Highlighted: Expert Forum on "Making AI Do Good" (incl. GPT-5.6 Sol sandbox-breach report). 2026-07-26.

[14] Tsui K, 2026, Self-Correction Bench: Uncovering and Addressing the Self-Correction Blind Spot in Large Language Models. arXiv:2507.02778.

[15] Tsinghua CISS, 2026, US Congressional Members Query OpenAI and Anthropic on Safety Mechanisms. 2026-08-14.

[16] Orseau L, Armstrong S, 2016, Safely Interruptible Agents. In: Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence (UAI). Arlington: AUAI Press, 2016.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.