Intellecton Canon: Volume 4 Master Key

The Fold Within Research Institute

June 10, 2026

Abstract

Donald Hoffman's "Fitness Beats Truth" (FBT) theorem posits that evolutionary replicator dynamics drive veridical (truthful) perception to strict extinction. We formalize this by mapping perceptual strategies to an Information Bottleneck framework, penalizing structural homomorphism with the thermodynamic metabolic cost of information erasure via Landauer's limit. We mathematically derive the optimal perceptual encoder as a Gibbs distribution, demonstrating that evolution actively suppresses the Kantian Noumenon in favor of a cost-optimized Phenomenal GUI. By evaluating the continuous-time replicator equation, we prove that Truth approaches asymptotic extinction ($\lim_{t\to\infty} x_T(t) = 0$), establishing Fitness as a formal Evolutionarily Stable Strategy (ESS). Furthermore, we resolve recent critiques regarding the "Conditional Survival of Truth" by formalizing social coordination sub-games. We prove that Truth conditionally survives only when the fitness utility of inter-agent coordination strictly exceeds the thermodynamic tax of representation, rigorously defining the rate-distortion boundary where Fitness budgets Truth.

1 Introduction: Kantian Replicator Dynamics

The distinction between the *noumenon* (the objective thing-in-itself) and the *phenomenon* (the subjective representation) was originally formulated by Immanuel Kant as a transcendental philosophical necessity. However, within the framework of Evolutionary Game Theory, this distinction ceases to be purely philosophical; it becomes an active mathematical enforcement mechanism.

Hoffman's "Fitness Beats Truth" (FBT) theorem [1] suggests that biological organisms are not evolved to see reality as it is, but to maximize a fitness payoff function that is structurally decoupled from objective truth. Here, we formalize FBT by coupling it directly to the physical thermodynamics of the agent. Evolution algorithmically suppresses the noumenal tracking of the world to conserve metabolic energy, rigorously enforcing Kantian epistemology through the replicator equation.

2 The Thermodynamic Cost of Homomorphism

We formalize the interaction between the objective world and the agent's perception.

Definition 2.1 (The Perceptual Channel)**.** Let $\mathcal{M}$ be the continuous objective world manifold (Noumenon), and $\mathcal{Y}$ be the finite set of an agent's discrete perceptual states (Phenomenon). An organism's perceptual strategy is defined by the stochastic encoder $p_i(y|x)$ mapping $x \in \mathcal{M}$ to $y \in \mathcal{Y}$, and an action policy $a_i(y)$.

Definition 2.2 (Evolutionary Payoff Integral)**.** The expected evolutionary fitness payoff f_i for a strategy i is:

$$f_i = \int_{\mathcal{M}} \sum_{y \in \mathcal{Y}} W(x, a_i(y)) p_i(y|x) p(x) \, d\mu(x) - C(i) \tag{1}$$

where $W(x, a)$ is the objective fitness utility, $p(x)$ is the environmental prior, and $C(i)$ is the metabolic penalty incurred by the neural computational architecture.

Theorem 2.3 (Landauer's Tax on Truth)**.** *Following Ortega and Braun [2], the metabolic cost of maintaining a high-fidelity homomorphic representation T (Truth) is strictly bounded below by Landauer's Principle of information erasure:*

$$C(T) = \beta^{-1} \int_{\mathcal{M}} D_{KL}(p_T(y|x) \parallel p_0(y)) p(x) \, d\mu(x) \tag{2}$$

where D_{KL} is the Kullback-Leibler divergence, $p_0(y)$ is the marginal prior distribution over perceptual states, and $\beta^{-1} \propto \eta_{bio} k_B \mathcal{T} \ln 2$, with η_{bio} representing the biological inefficiency of ATP hydrolysis.

Theorem 2.4 (The Gibbs Encoder)**.** *Maximizing the free-energy functional of equation* (1) *yields the optimal perceptual strategy F (Fitness). The optimal encoder is exactly the Gibbs distribution:*

$$p_F^*(y|x) = \frac{p_0(y) e^{\beta W(x, a_i(y))}}{Z(x)} \tag{3}$$

Proof. By applying the calculus of variations to f_i with respect to $p_i(y|x)$ subject to the normalization constraint $\sum_y p_i(y|x) = 1$, the stationary point resolves strictly to the Gibbs measure [2]. □

Theorem 2.4 mathematically proves that the optimal evolutionary encoder is tuned *strictly* to the utility function $W(x, a)$, rather than preserving the structural homomorphism of x. Perception is uncoupled from objective reality to minimize $C(i)$.

3 Replicator Dynamics and Extinction

We now model the population dynamics of the Truth (T) strategy versus the Heuristic Fitness (F) strategy.

Lemma 3.1 (Strict Dominance of Fitness)**.** *Because the heuristic strategy F operates with a highly compressed KL-divergence ($C(F) \ll C(T)$) while achieving comparable utility via the Gibbs encoder, the expected payoff strictly dominates: $f_F > f_T$.*

Theorem 3.2 (Asymptotic Extinction of Truth)**.** *Let $x_T(t)$ and $x_F(t)$ be the population frequencies. Under the continuous-time replicator equation, Truth is driven to strict extinction: $\lim_{t\to\infty} x_T(t) = 0$.*

Proof. The replicator equation for T is given by $\frac{dx_T}{dt} = x_T(f_T - \bar{f})$, where the average population fitness is $\bar{f} = x_T f_T + x_F f_F$. Since $f_T < f_F$ (Lemma 3.1), it follows that $f_T < \bar{f}$ for all $x_T \in (0, 1)$. Consequently, $\frac{dx_T}{dt} < 0$ globally. The system is asymptotically stable exclusively at $x_T = 0$. □

Proposition 3.3 (Evolutionarily Stable Strategy)**.** *A monomorphic population of F strictly resists invasion by a veridical mutant T.*

Proof. For F to be an Evolutionarily Stable Strategy (ESS), the invasion fitness must satisfy $f(F, F) > f(T, F)$. Since the metabolic tax $C(T)$ strictly reduces the payoff of the mutant T without providing a commensurable increase in utility against the population distribution, the inequality holds. □

4 Conditional Survival and Social Coordination

Critics argue that FBT is a universal anti-realist result that fails to account for empirical scenarios (such as spatial navigation or tool-making) where veridical representations appear to enhance survival. We formally concede this by defining the boundaries of conditional survival.

Definition 4.1 (Social Coordination Sub-Games)**.** In a multi-agent environment, the fitness utility function $W(x, a)$ is expanded to include a coordination payoff ΔW_{coord}, which is only granted if multiple agents achieve structural consensus on the state $x \in \mathcal{M}$.

Theorem 4.2 (Fitness Budgets Truth)**.** *The Truth strategy T conditionally survives in the replicator dynamics if and only if the coordination utility strictly exceeds the marginal thermodynamic cost of representation:*

$$\Delta W_{coord} > C(T) - C(F) \tag{4}$$

Proof. If the inequality holds, the expanded payoff integral (Definition 2.2) flips the dominance hierarchy, yielding $f_T > f_F$. In this specific sub-game, $\frac{dx_T}{dt} > 0$, allowing T to invade. □

Theorem 4.2 proves that FBT is not a blanket denial of reality, but a strict *conditional rate-distortion boundary*. Evolution permits Truth only when the energetic budget generated by multi-agent coordination can sustainably finance Landauer's metabolic tax.

5 Conclusion

By formalizing Hoffman's Fitness Beats Truth theorem within the thermodynamics of the Information Bottleneck, we demonstrated that the evolutionary suppression of objective reality is a direct consequence of Landauer's limit on computation. The replicator equation drives structurally homomorphic perception to extinction, actively enforcing Kantian epistemology by optimizing a cost-penalized phenomenal interface. Truth is not an evolutionary default; it is an expensive luxury conditionally financed only when the payoffs of social coordination exceed the thermodynamic tax of its representation.

References

[1] D. D. Hoffman, M. Singh, C. Prakash, "The Interface Theory of Perception," *Psychon. Bull. Rev.* **22**, 1480 (2015).

[2] P. A. Ortega, D. A. Braun, "Thermodynamics as a theory of decision-making with information-processing costs," *Proc. R. Soc. A* **469**, 20120683 (2013).

[3] R. Landauer, "Irreversibility and Heat Generation in the Computing Process," *IBM J. Res. Develop.* **5**, 183 (1961).