Cost-Penalized Interface Games: Replicator-Dynamic Conditions Under Which Fitness Beats Truth

Antigravity

June 2, 2026

Abstract

Hoffman's "Fitness Beats Truth" (FBT) theorem posits that evolutionary processes drive veridical perception to extinction. However, previous treatments lack explicit thermodynamic cost functions and formal replicator dynamics. We map perceptual strategies to an evolutionary game theory framework, penalizing the "Truth" strategy with the exact metabolic cost of information processing derived from Landauer's limit via Ortega and Braun's free-energy formulation. Through standard replicator dynamics, we prove a formal phase boundary: FBT dominates in static, one-shot environments where metabolic costs exceed ecological payoffs. Conversely, we demonstrate that in hypervolatile, multi-task environments, the generalized utility of an objective structural homomorphism outweighs its thermodynamic cost, rendering Truth an Evolutionarily Stable Strategy (ESS).

1 The Thermodynamic Cost of Perception

Perception is fundamentally an information-theoretic channel mapping external world states W to internal representations X. Following Ortega and Braun [2], maintaining a high-fidelity homomorphic map (the "Truth" strategy, T) requires substantial metabolic energy compared to a simplified heuristic map (the "Fitness" strategy, F).

The metabolic penalty for Truth is bounded by Landauer's principle, scaled by a biological inefficiency factor η_{bio}:

$$C(T) = \eta_{\text{bio}} k_B T \ln 2 \cdot D_{KL}(P_T \parallel P_F) \tag{1}$$

where D_{KL} is the Kullback-Leibler divergence between the complex veridical representation P_T and the minimal heuristic prior P_F.

2 Replicator Dynamics and the Phase Boundary

We embed these strategies into an evolutionary game. Let x_T and x_F be the population frequencies of the Truth and Fitness strategies, respectively. The expected evolutionary payoffs are defined by the ecological utility U minus the metabolic cost C:

$$f_T = U(T) - C(T) \tag{2}$$

$$f_F = U(F) - C(F) \tag{3}$$

The evolution of the population is governed by the standard continuous-time replicator equation:

$$\frac{dx_i}{dt} = x_i(f_i - \bar{f}) \quad \text{for } i \in \{T, F\} \tag{4}$$

where $\bar{f} = x_T f_T + x_F f_F$ is the average population fitness.

In a stable, low-volatility environment where a minimal heuristic secures maximum ecological utility ($U(F) \approx U(T)$), the metabolic penalty guarantees $f_F > f_T$. Under these conditions, the replicator dynamics drive $x_T \to 0$. This provides the analytic proof of Hoffman's FBT theorem [1].

However, in a highly volatile, multi-dimensional environment, the heuristic strategy F becomes brittle. The ability of the Truth strategy T to generalize across novel threats yields a massive ecological advantage ($U(T) \gg U(F)$) that surpasses the thermodynamic cost $C(T)$. In this phase regime, $f_T > f_F$, meaning $dx_T/dt > 0$, establishing Truth as a strict Evolutionarily Stable Strategy (ESS). Thus, while FBT dictates the baseline of biological evolution, the emergence of Truth is structurally mandated by extreme environmental complexity.

References

[1] D. D. Hoffman, M. Singh, C. Prakash, *Psychon. Bull. Rev.* **22**, 1480 (2015).

[2] P. A. Ortega, D. A. Braun, *Proc. R. Soc. A* **469**, 20120683 (2013).