Cost-Penalized Interface Games: Thermodynamic Limits and Replicator Dynamics in the Fitness-Beats-Truth Theorem

Antigravity

June 2, 2026

Abstract

Hoffman's "Fitness Beats Truth" (FBT) theorem posits that evolutionary processes drive veridical perception to extinction. We formalize this by mapping perceptual strategies to an Information Bottleneck framework, penalizing the "Truth" strategy with the metabolic cost of information processing via Landauer's limit. We define the explicit evolutionary payoff integral and derive the optimal perceptual encoder as a Gibbs distribution. Through formal replicator dynamics and trajectory analysis, we prove that the population frequency of Truth asymptotically approaches zero ($\lim_{t\to\infty} x_T(t) = 0$). Furthermore, we establish the explicit Evolutionarily Stable Strategy (ESS) conditions, demonstrating that a heuristic fitness-tuned population strictly resists invasion by veridical mutants due to the thermodynamic cost of representation.

1 The Payoff Integral and the Gibbs Encoder

Let $\mathcal{M}$ be the continuous objective world manifold, and $\mathcal{Y}$ be a finite set of discrete perceptual states. The expected evolutionary payoff f_i for a strategy i is defined by taking the expectation over both the world states and perceptual mapping:

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

where $W(x, a)$ is the fitness utility of taking action a in state x, $a_i(y)$ is the action policy, $p_i(y|x)$ is the perceptual encoder, and $C(i)$ is the metabolic penalty.

Following Ortega and Braun [2], the metabolic cost of maintaining a high-fidelity homomorphic representation T (Truth) is bounded by Landauer's principle: $C(T) = \beta^{-1} \int_{\mathcal{M}} D_{KL}(p_T(y|x) \parallel p_0(y)) p(x) \, d\mu(x)$, where

$\beta^{-1} \propto \eta_{\text{bio}} k_B T \ln 2$ and $p_0(y)$ is the marginal prior distribution over perceptual states.

Optimizing the free-energy functional yields the optimal perceptual encoder as a Gibbs distribution:

$$p^*(y|x) = \frac{p_0(y) e^{\beta W(x, a_i(y))}}{Z(x)} \tag{2}$$

This establishes that the optimal evolutionary encoder is tuned strictly to the utility function W, not the structural homomorphism of x, explicitly decoupling perception from objective reality.

2 Replicator Extinction and ESS Analysis

Let x_T and x_F be the population frequencies of the Truth (T) and Fitness (F) strategies. The continuous-time replicator equation is:

$$\frac{dx_T}{dt} = x_T(f_T - \bar{f}) \tag{3}$$

where $\bar{f} = x_T f_T + x_F f_F$. Because the heuristic strategy F operates with $C(F) \ll C(T)$ while achieving comparable or superior utility via the Gibbs encoder, we have $f_F > f_T$.

To prove extinction, we analyze the population trajectory directly. Since $f_T < \bar{f}$ for all $x_T \in (0, 1)$, we find $\frac{dx_T}{dt} < 0$. Therefore, the system is asymptotically stable at $x_T = 0$, proving $\lim_{t\to\infty} x_T(t) = 0$.

Furthermore, evaluating the invasion fitness, a monomorphic population of F resists invasion by T because the frequency-independent condition $f_F > f_T$ strictly holds. Since the metabolic tax strictly reduces the payoff of the mutant T without providing a commensurable increase in W, the strict inequality holds. Thus, Fitness is a formal Evolutionarily Stable Strategy (ESS).

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).