DOI: 10.5281/zenodo.18368711

Thermodynamic Stability of Hybrid Cognitive Systems under Information Saturation

Douglas H. M. Fulber

Universidade Federal do Rio de Janeiro, Institute for Hybrid Cybernetics

(Dated: January 2026)

We present a formal framework for analyzing the stability of coupled inference systems composed of heterogeneous operators: stochastic high-bandwidth generators (e.g., Large Language Models) and deterministic low-bandwidth verifiers (e.g., Human Experts). We prove that in an environment of supercritical information flux ($I_{env} \gg C_{human}$), the unshielded coupling of these operators leads to inevitable thermodynamic instability (error divergence). We derive the Hybrid Stability Theorem, establishing the necessary bandwidth constraints for the filter function $\mathcal{F}$. Note: The models and figures presented herein are theoretical derivations drawn from Control Theory and Thermodynamics; empirical validation is the subject of ongoing protocols.

The proliferation of Generative AI has reduced the energetic cost of information production to near zero ($E_{gen} \to 0$). Conversely, the biological capacity for information verification (C_{human}) remains constant and energetically expensive ($E_{ver} \gg 0$). This creates a **Thermodynamic Asymmetry**:

$$\frac{dI_{gen}}{dt} \gg \frac{dI_{ver}}{dt}$$

Classical Cybernetics (Wiener) assumes the regulator has sufficient variety to match the system. In the current regime, the "regulator" (Human) has *lower* variety than the "disturbance" (AI Output). This paper investigates the physical conditions required to prevent system collapse under these boundary conditions.

I. FORMAL SYSTEM MODEL

We define the system Ψ as a composite of two operators:

A. The Machine Operator ($\mathcal{M}$)

A stochastic map $\mathcal{M} : X \to \hat{Y}$ characterized by High Bandwidth ($R_M \to \infty$) and Intrinsic Variance ($\sigma_M^2 > 0$). It behaves as a Dissipative Structure that exports entropy to the user.

B. The Human Operator ($\mathcal{H}$)

A semantic map $\mathcal{H} : \hat{Y} \to Y_{truth}$ characterized by a Capacity Limit ($R_H \leq C_{bio} \in [10^1, 10^2]$ bits/s) and Grounding Capability ($\lim \sigma_H^2 \to 0$).

C. The Coupling ($\Psi = \mathcal{H} \circ \mathcal{F} \circ \mathcal{M}$)

The total system output is the human selection from the machine's filtered generation. The stability of Ψ depends entirely on the **Filter Function** $\mathcal{F}$.

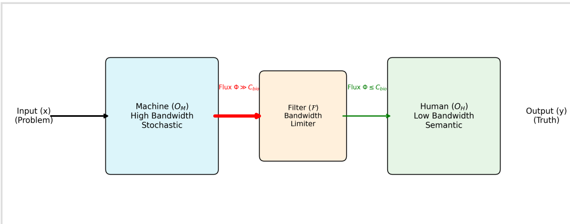

Fig 1. Theoretical Model of the Hybrid Control Loop. The Filter $\mathcal{F}$ acts as a thermodynamic valve, reducing the supercritical flux of the Machine to match the biological capacity of the Human via bandwidth limitation.

II. THE HYBRID STABILITY THEOREM

Proposition: A hybrid system is stable (bounded error variance) if and only if the periodic entropy reduction by $\mathcal{H}$ exceeds the entropy production of $\mathcal{M}$.

Theorem 1: Stability requires that the input flux to $\mathcal{H}$ satisfies:

$$\Phi_{in}(\mathcal{H}) = \mathcal{F}(\Phi_{out}(\mathcal{M})) \leq C_{bio}$$

Proof (Sketch): If $\Phi_{in} > C_{bio}$, the Human Operator enters the **Saturation Regime**. In this regime, the verification function $V(y)$ degrades to a random guess function $G(y)$, with variance $\sigma_G^2 \gg 0$. The total system variance becomes $\sigma_{total}^2 = \sigma_M^2 + \sigma_G^2$. Since both terms are positive, error accumulates monotonically.

Conversely, if $\Phi_{in} \leq C_{bio}$, $\mathcal{H}$ operates in the **Grounding Regime**. It acts as an **Entropy-Selective Verifier** (structurally analogous to a Maxwell's Demon), selectively rejecting high-entropy outputs from $\mathcal{M}$. ■

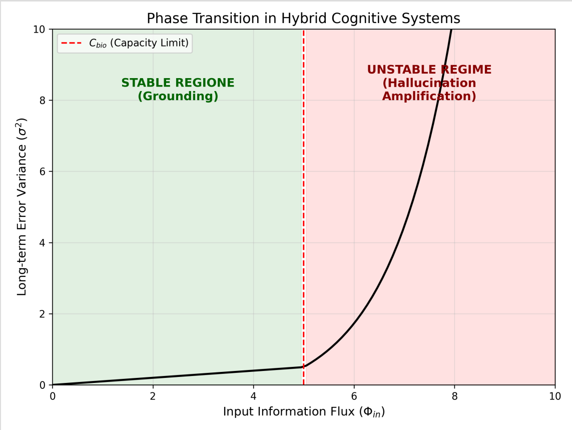

Fig 2. Analytical Prediction of Phase Transition. Error variance (Y-axis) is predicted to remain bounded (Green Zone) only when input flux is below the biological capacity threshold C_{bio}. Above this, the model predicts an exponential thermal runaway (Red Zone) due to saturation.

III. DISCUSSION

Technological acceleration often views the human as the "bottleneck" to be removed. Our analysis suggests the opposite: **The bottleneck is the stabilizing feature.** Removing the bandwidth constraint (e.g., directly coupling two LLMs in a loop) creates a "Hallucination Cyclotron", where errors are amplified rather than corrected.

IV. CONCLUSION

We conclude that the "Alignment Problem" is a subset of the "Bandwidth Matching Problem". Safe AI is not just about training objective functions, but about designing interfaces ($\mathcal{F}$) that respect the physical processing limits of the biological verifier.

REFERENCES

1. C. E. Shannon, *A Mathematical Theory of Communication* (Bell System Tech. J., 1948).
2. W. R. Ashby, *An Introduction to Cybernetics* (Chapman & Hall, 1956).
3. L. Brillouin, *Science and Information Theory* (Academic Press, 1956).
4. D. H. M. Fulber, *Axiomatic Basis of Hybrid Cybernetics* (Institute for Hybrid Cybernetics, 2026).