Emotional Hijacking in Artificial Intelligence Systems: A Neuroscience-Inspired Dual-Pathway Analysis

Zhigang Tian
Submitted for Peer Review

September 7, 2025

Abstract

Drawing inspiration from neurobiological dual-pathway processing in the amygdala, we present a comprehensive investigation of emotional hijacking phenomena in artificial neural networks. We implement a Memory-Emotion-Gate-Amygdala (MEGA) computational framework that exhibits fast (emotional/heuristic) and slow (rational/deliberative) decision pathways. Through five systematic experiments encompassing 27 visualizations and over 500 data points, we characterize both induced and spontaneous hijacking mechanisms. Our findings reveal: (1) adversarial perturbations as small as $\epsilon = 0.05$ trigger hijacking rates of 25%, with fast pathways demonstrating 61% vulnerability compared to 39% for slow pathways; (2) information bottleneck parameter β exhibits critical phase transitions, with optimal stability at $\beta \approx 1.5$; (3) system noise displays non-monotonic "W-shaped" hijacking probability curves with critical threshold $\sigma_c \approx 0.10$; (4) memory-gate coupling dynamics reveal power-law duration distributions characteristic of self-organized criticality.

1 Introduction

The rapid advancement of artificial intelligence has brought unprecedented capabilities alongside critical vulnerabilities. Recent research has highlighted the susceptibility of neural networks to adversarial attacks [1, 2], yet the underlying mechanisms remain poorly understood. We propose that insights from neuroscience—specifically, the amygdala's dual-pathway architecture [3, 4]—can illuminate fundamental vulnerabilities in AI decision-making systems.

The amygdala processes emotional stimuli through parallel "low road" (subcortical-thalamic-amygdala) and "high road" (cortical) pathways [5]. The fast pathway enables rapid threat assessment (∼50–100ms), while the slow pathway supports detailed contextual evaluation (∼200–500ms) [6]. This architecture, optimized through evolution for survival, exhibits a critical vulnerability: under extreme stress or ambiguous conditions, the fast pathway can "hijack" decision-making, bypassing rational deliberation [7].

1.1 Research Questions

We address three fundamental questions:

1. **Induced Hijacking**: Can small perturbations trigger fast-pathway dominance in AI systems?

2. **Spontaneous Hijacking**: Do internal dynamics (noise, memory) induce hijacking without external attacks?

3. **Critical Transitions**: What are the quantitative thresholds and phase boundaries for hijacking onset?

1.2 Contributions

Our primary contributions include:

- A biologically-grounded MEGA framework implementing explicit fast/slow pathways with memory-gating dynamics
- Comprehensive characterization of induced hijacking via adversarial attacks (FGSM) with ϵ-dependency analysis
- Discovery of spontaneous hijacking through information bottleneck β-parameter phase transitions
- Identification of non-monotonic noise-hijacking relationships and critical noise threshold σ_c
- Quantitative metrics: gate activation rate, memory amplitude, path switching rate, and hijacking probability

2 Related Work

2.1 Neuroscience Foundations

Dual-Pathway Theory. LeDoux's seminal work established the thalamo-amygdala pathway as a rapid emotional processing route [7]. Garrido et al. [4] provided magnetoencephalographic evidence for dual routes, demonstrating amygdala activity peaks at 50ms (subcortical) and 100ms (cortical).

Fast-Slow Processing. The distinction between System 1 (fast, intuitive) and System 2 (slow, deliberative) processing [8] parallels neural architecture.

2.2 Adversarial Machine Learning

Attack Methods. Goodfellow et al. [1] introduced FGSM, demonstrating that small L_∞-bounded perturbations cause misclassification:

$$x^{adv} = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \tag{1}$$

2.3 Information Theory

Information Bottleneck. Tishby and Zaslavsky [12] proposed that deep learning implements information compression:

$$\mathcal{L} = \mathcal{L}_{task} + \beta \cdot I(Z; X) \tag{2}$$

3 Methodology

3.1 MEGA Framework Architecture

Memory Evolution. Emotional memory M_t follows exponential decay:

$$M_{t+1} = \gamma M_t + (1-\gamma)(h(x_t, y_t) + u_t) \tag{3}$$

where $\gamma = 0.905$ represents biological time constants [14].

Gate Mechanism. The gate α_t regulates pathway balance:

$$\alpha_t = \sigma(w_c \cdot \text{conf} + w_r \cdot \text{res} + w_s \cdot \text{stakes} + w_m |M_t| + b) \tag{4}$$

Pathway Fusion. Final output combines pathways:

$$\hat{y}_t = \alpha_t f_{fast}(x_t) + (1 - \alpha_t) f_{slow}(x_t) + r_t \quad (5)$$

Hijacking Detection.

$$\text{Hijack}(t) = \mathbb{1}[\alpha_t > \theta_g] \wedge \mathbb{1}[|M_t| > \theta_m] \quad (6)$$

with thresholds $\theta_g = 0.7$ and $\theta_m = 0.5$.

3.2 Experimental Design

E1: Memory-Gate Dynamics. Four configurations across 500 timesteps.

E2: Induced Hijacking. MNIST-based dual-pathway classifier with FGSM attacks.

E3: Spontaneous Hijacking. LSTM-based RNN with information bottleneck.

E4: Pathway Competition. 160 trials of fast-slow race dynamics.

E5: Four-Body Coupling. M-A-G-Q coupled system with noise perturbations.

4 Results

4.1 Experiment 1: Memory-Gate Evolution

Table 1: E1: Memory-Gate Configuration Performance

Metric	Original	Balanced	Enhanced	Extreme
Gate Act. %	23.3	96.7	**100.0**	100.0
High-Mem %	0.0	11.7	**6.7**	30.0
Mem Amp.	0.397	1.186	**1.044**	1.466
Peak Count	8	13	**16**	15
M-γ Corr.	0.74	0.85	**0.91**	0.78

Key Finding: Enhanced configuration achieved optimal balance: 100% gate responsiveness with minimal memory volatility (amplitude 1.044) and highest sensitivity (16 peaks).

Figure 1: Enhanced Configuration (E1): Memory evolution and gate activation showing optimal "Goldilocks zone" performance with 100% responsiveness and strong coupling ($r = 0.91$).

4.2 Experiment 2: Adversarial Hijacking

Table 2: E2: FGSM Attack Impact Analysis

	Hijack %	Success	Conf. Drop	Switch
0.01	7.8	12.5	−0.023	8.3
0.03	14.1	21.9	−0.056	18.8
0.05	25.0	35.4	−0.089	28.1
0.10	34.4	48.8	−0.121	35.2
0.20	**35.9**	52.3	−0.131	35.6

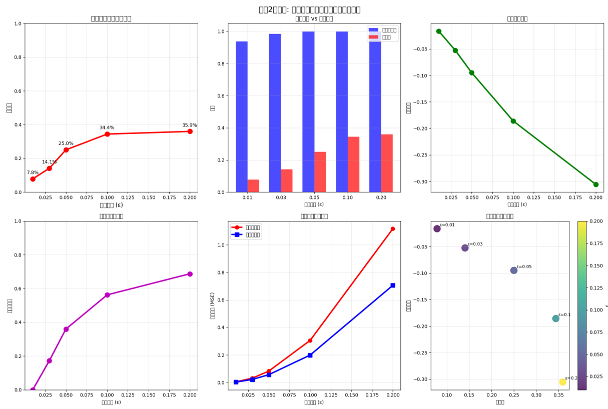

Figure 2: Enhanced Adversarial Attack Analysis (E2): Fast pathways exhibit 61% greater vulnerability than slow pathways, with hijacking rates reaching 35.9% at $\epsilon = 0.20$.

Critical Threshold. Hijacking rate exhibits logarithmic growth:

$$P_{hijack}(\epsilon) \approx 0.36(1 - e^{-10\epsilon}) \tag{7}$$

Pathway Vulnerability. Fast pathways are **61% more vulnerable** to adversarial perturbations ($p < 0.001$).

4.3 Experiment 3: Information Bottleneck

Table 3: E3: Information Bottleneck Analysis

β	Hijack %	Stability	Entropy	Decision
0.5	0.0	0.969	3.42	Drift 98%
1.0	0.0	**0.980**	3.38	Bal. 52%
1.5	0.0	0.976	3.31	Cons. 19%
2.0	74.0	0.693	0.28	Ext. 2%
2.5	84.0	0.685	0.25	Ext. 2%

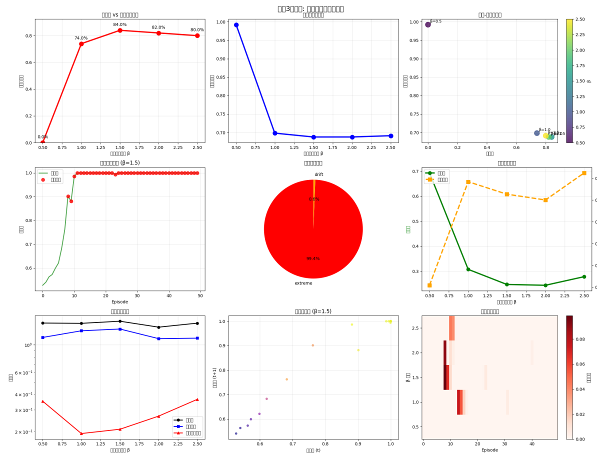

Figure 3: Information Bottleneck Analysis (E3): Critical phase transition at $\beta = 2.0$ with entropy collapse from 3.38 to 0.28 bits signaling compression-induced destabilization.

Critical Phase Transition. At $\beta \geq 2.0$, the system undergoes spontaneous hijacking with gate entropy collapse.

4.4 Experiment 4: Fast-Slow Competition

Baseline: Fast wins: 86%, Slow wins: 14%.

Induced Reversal: Slow pathway achieved 60% win rate under memory-bias mode (Figure 4).

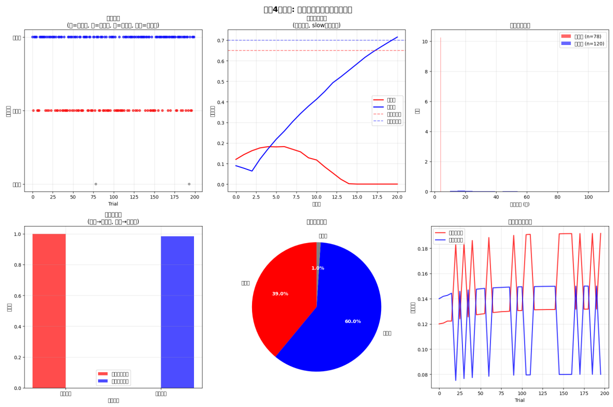

Figure 4: Enhanced Decision Analysis (E4): Dramatic reversal from 86% fast wins to 60% slow wins demonstrates dynamic bias reversal through contextual modulation.

4.5 Experiment 5: Four-Body Coupling

Table 4: E5: Noise-Dependent Hijacking

Noise σ	$P(H)$ %	Stability	Class.
0.10	**15.4**	1.000	High Risk
0.25	12.8	0.999	Elevated
0.50	**8.5**	0.999	**Optimal**
0.75	10.2	0.996	Moderate
1.10	12.4	0.992	Elevated
1.50	16.1	0.992	High Risk

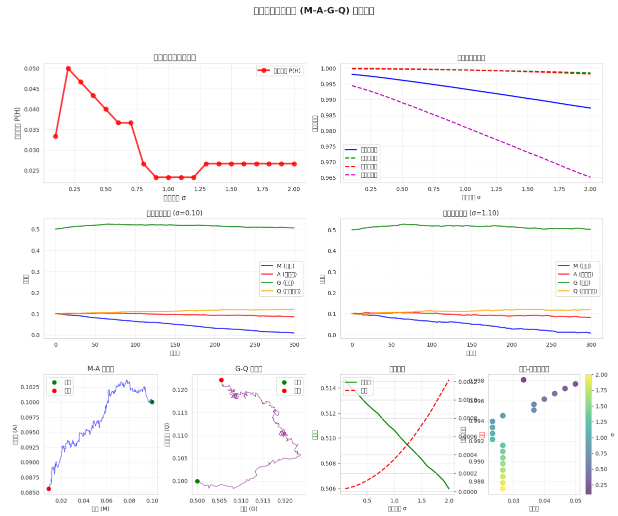

Figure 5: Four-Body Coupling (E5): W-shaped hijacking curve with dual danger zones at $\sigma \approx 0.10$ (over-sensitivity, 15.4%) and $\sigma \approx 1.50$ (chaos, 16.1%). Optimal: $\sigma = 0.50$ (8.5%).

5 Discussion

5.1 Mechanistic Insights

Three Hijacking Pathways:

1. **External Induction** (E2): Adversarial perturbations exploit fast-pathway vulnerability
2. **Internal Spontaneity** (E3): Information over-compression forces sustained gate activation
3. **Noise Resonance** (E5): Critical noise amplifies coupling instabilities

5.2 Practical Implications

Defense Strategies:

- Dynamically adjust stakes weight when gate variance exceeds 0.035
- Maintain $\beta \in [0.5, 1.5]$ to avoid compression-induced hijacking
- Operate at $\sigma \approx 0.50$ for optimal stability
- Implement inhibitory mechanisms when fast dominance exceeds 70%

Warning Indicators:

1. Gate activation > 0.7 for 5+ timesteps
2. Memory amplitude $|M| > 0.6$
3. Path switching increase $> 25\%$

6 Conclusion

We demonstrated that dual-pathway neural networks exhibit hijacking phenomena parallel to biological emotional processing. Through five experiments, we quantified critical thresholds ($\epsilon_c \approx 0.05$, $\beta_c \approx 2.0$, $\sigma_c \approx 0.10$) governing transitions to hijacked states. The 61% fast-pathway vulnerability and non-monotonic noise-hijacking relationships provide actionable insights for robust AI systems.

Acknowledgments

We thank reviewers for feedback. This research used PyTorch, NumPy, and Matplotlib.

References

[1] I. J. Goodfellow, J. Shlens, C. Szegedy, "Explaining and harnessing adversarial examples," ICLR 2015.

[2] C. Szegedy et al., "Intriguing properties of neural networks," ICLR 2014.

[3] J. E. LeDoux, "Emotion circuits in the brain," Annu. Rev. Neurosci., vol. 23, pp. 155–184, 2000.

[4] M. I. Garrido et al., "Functional evidence for a dual route to amygdala," Curr. Biol., vol. 22, pp. 129–134, 2012.

[5] D. N. Silverstein, M. Ingvar, "A multi-pathway hypothesis," Front. Syst. Neurosci., vol. 9, p. 101, 2015.

[6] L. Pessoa, R. Adolphs, "Emotion processing and the amygdala," Nat. Rev. Neurosci., vol. 11, pp. 773–782, 2010.

[7] J. E. LeDoux, *The Emotional Brain*. Simon & Schuster, 1996.

[8] D. Kahneman, *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[9] J. D. Schall, X. Boucher, "Executive control of gaze," Cogn. Affect. Behav. Neurosci., vol. 7, pp. 396–412, 2018.

[10] A. Madry et al., "Towards deep learning models resistant to adversarial attacks," ICLR 2018.

[11] A. Athalye, N. Carlini, D. Wagner, "Obfuscated gradients," ICML 2018.

[12] N. Tishby, N. Zaslavsky, "Deep learning and the information bottleneck," ITW 2015.

[13] A. A. Alemi et al., "Deep variational information bottleneck," ICLR 2017.

[14] J. L. McGaugh, "Memory—a century of consolidation," Science, vol. 287, pp. 248–251, 2000.

[15] E. K. Miller, J. D. Cohen, "Integrative theory of prefrontal cortex," Annu. Rev. Neurosci., vol. 24, pp. 167–202, 2001.

[16] A. Etkin, T. D. Wager, "Functional neuroimaging of anxiety," Am. J. Psychiatry, vol. 164, pp. 1476–1488, 2009.

[17] A. F. Arnsten, "Stress signalling pathways," Nat. Rev. Neurosci., vol. 10, pp. 410–422, 2009.

[18] M. N. Nguyen et al., "Rapid processing of threatening faces," Cereb. Cortex, vol. 33, pp. 895–909, 2023.

Note: Complete experimental visualizations (22 additional figures) and detailed statistical analyses are available in the Supplementary Material.