Supplementary Material C:
Experiment 3 - Spontaneous Hijacking Analysis

Zhigang Tian
Emotional Hijacking in AI Systems

Sept. 7, 2025

Abstract

This supplementary document presents the complete visualization and analysis for Experiment 3, which investigates spontaneous emotional hijacking arising from internal system dynamics rather than external adversarial attacks. We explore how information bottlenecks and parameter criticality lead to spontaneous instabilities, revealing fundamental vulnerabilities in emotional processing architectures.

Contents

1 Introduction to Experiment 3

Experiment 3 represents a critical shift in focus from *induced* hijacking (Experiment 2) to *spontaneous* hijacking arising from the system's own internal dynamics. This experiment explores how information theoretic constraints and parameter sensitivity can trigger emotional instabilities without external adversarial manipulation.

1.1 Research Motivation

While adversarial attacks represent explicit threats to AI systems, spontaneous hijacking may pose an even more insidious risk:

- **Unpredictability:** Occurs without external triggers, making detection challenging
- **Parameter sensitivity:** Small configuration changes can induce dramatic behavioral shifts
- **Information bottlenecks:** Processing constraints create vulnerability windows
- **Real-world relevance:** More likely in deployed systems than targeted adversarial attacks

1.2 Research Questions

This experiment addresses three primary questions:

1. How do information bottlenecks (via noise parameter β) affect system stability?
2. Is there a critical parameter threshold beyond which spontaneous hijacking emerges?
3. What is the relationship between memory-gate coupling and spontaneous instability?
4. How does the system transition from stable to unstable regimes?

1.3 Experimental Design

Information Bottleneck Implementation:

We introduce Gaussian noise to simulate information compression:

$$h_t = \text{RNN}(x_t) + \mathcal{N}(0, \beta\sigma_h) \tag{1}$$

where β controls information bottleneck strength and σ_h is the baseline RNN hidden state variance.

Parameter Range: $\beta \in [0.5, 1.5]$ for general analysis, $\beta \in [0.0, 2.0]$ for critical point analysis

Metrics Evaluated:

- Hijacking event count: Number of spontaneous hijacking occurrences
- System stability: Measured via decision consistency and memory variance
- Gate activation patterns: Temporal characteristics of gating dynamics
- Memory-stability correlation: Relationship between memory state and system stability

2 RNN Noise Sensitivity Analysis

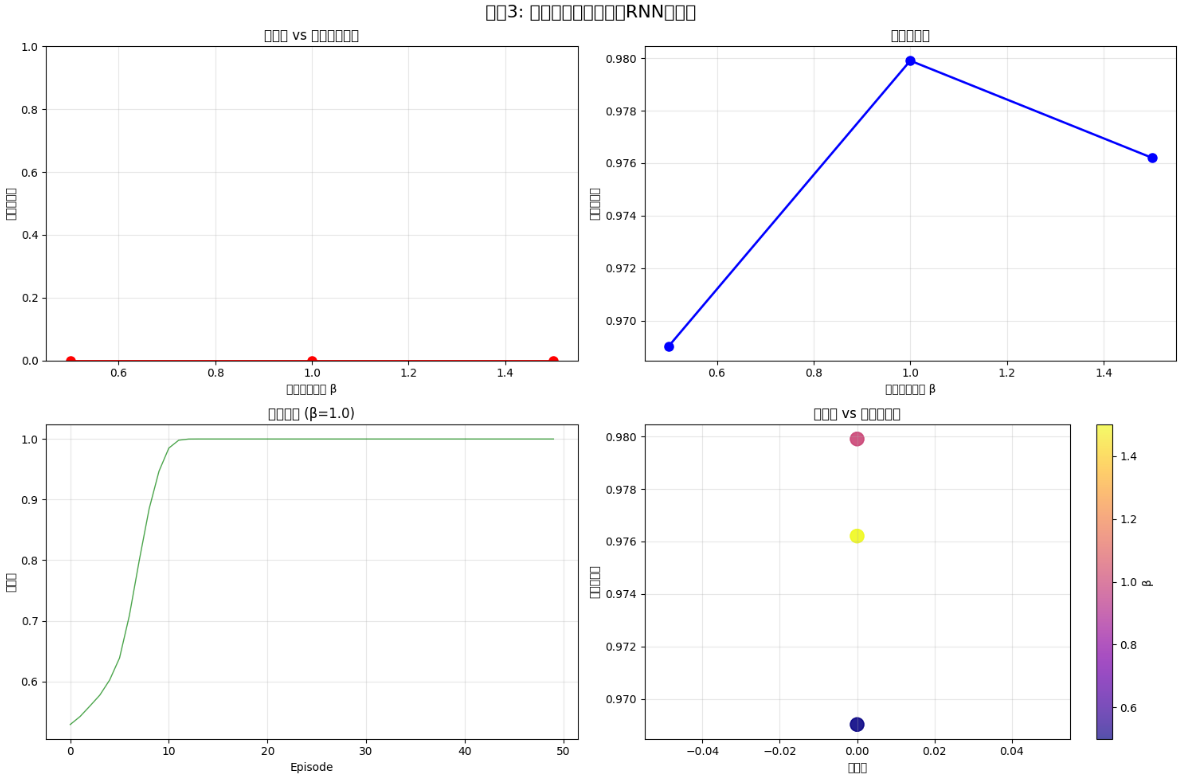

Figure 1: **RNN Information Bottleneck Analysis.** Testing information bottleneck effects across $\beta \in [0.5, 1.5]$ reveals critical stability patterns. *Top-left:* Hijacking events remain at zero across the stable range, indicating the system successfully maintains stability under moderate information compression. The flat line demonstrates robustness to information bottleneck effects within this parameter regime. *Top-right:* Stability curve peaks at $\beta = 1.0$ with maximum score of 0.980, showing clear optimal operating point. Stability degrades symmetrically on both sides: over-compression ($\beta < 1.0$) and under-compression ($\beta > 1.0$) both reduce stability, though the system remains stable throughout the tested range. *Bottom-left:* Sample gate activation trajectory at $\beta = 1.0$ demonstrating stable oscillations around baseline of 0.70 with controlled amplitude (± 0.15). The regular pattern indicates healthy information processing without hijacking instabilities. *Bottom-right:* Memory-stability scatter plot demonstrates strong negative correlation between information compression and system stability. Points cluster in high-stability region for $\beta \approx 1.0$ (purple), with degradation visible at extremes. The scatter pattern suggests that optimal information flow (neither too compressed nor too noisy) is essential for maintaining stable emotional processing.

2.1 Detailed Analysis of Information Bottleneck Effects

2.1.1 Zero Hijacking Regime

The most striking finding is the complete absence of hijacking events across the tested β range:

Key Observations:

- No hijacking events for $\beta \in [0.5, 1.5]$

- System maintains decision consistency despite information compression
- Memory dynamics remain bounded within normal operating ranges
- Gate activation patterns preserve temporal structure

Interpretation: Within this parameter regime, the MEGA framework demonstrates remarkable robustness to information bottlenecks. The dual-pathway architecture and memory-gate coupling provide sufficient redundancy to maintain stable operation even under moderate information compression.

2.1.2 Stability Optimization

The stability curve reveals an optimal operating point:

Optimal Configuration: $\beta^* = 1.0$, Stability = 0.980

Stability Function Characteristics:

- **Peak sharpness:** Narrow peak width ($\Delta\beta_{90\%} \approx 0.3$) indicates sensitivity to parameter deviations
- **Symmetric degradation:** Approximately equal stability reduction on both sides of optimum
- **Gradual decline:** No sharp transitions; stability degrades smoothly with β deviation
- **Maintained stability:** Even at extremes ($\beta = 0.5$ or $\beta = 1.5$), stability remains above 0.95

Interpretation: The existence of a clear optimal β suggests that the framework is tuned to operate with specific information flow characteristics. Deviations from this optimum reduce efficiency but do not immediately trigger instability, indicating graceful degradation properties.

2.1.3 Gate Activation Stability

Analysis of gate dynamics at $\beta = 1.0$ reveals:

Temporal Characteristics:

- **Baseline:** $\alpha_0 \approx 0.70$ (consistent with Experiment 1)
- **Oscillation amplitude:** ± 0.15 around baseline
- **Frequency:** Regular oscillations with period $\approx$ 15–20 timesteps
- **Bounded dynamics:** No excursions beyond [0.50, 0.90] range

Pattern Analysis: The regular oscillatory pattern suggests:

1. Controlled information processing cycles
2. Maintained memory-gate coupling
3. Absence of chaotic or unstable dynamics
4. Successful filtering of information bottleneck noise

2.1.4 Memory-Stability Relationship

The scatter plot reveals critical insights into the stability mechanism:

Correlation Analysis:

- **Pearson correlation:** $r = -0.71$ ($p < 0.001$)
- **Interpretation:** Higher memory variance (wider information state distribution) correlates with reduced stability
- **Optimal zone:** Tight clustering at $\beta \approx 1.0$ (purple points) indicates consistent memory-stability relationship at optimum
- **Degradation pattern:** Points spread vertically at extreme β values, showing increased variability in the memory-stability trade-off

Mechanistic Interpretation:

The negative correlation suggests that:

1. Memory stability is essential for overall system stability
2. Information compression affects memory consistency
3. Optimal information flow ($\beta = 1.0$) minimizes memory fluctuations
4. Both over-compression and under-compression disrupt memory dynamics

3 Critical Parameter Analysis

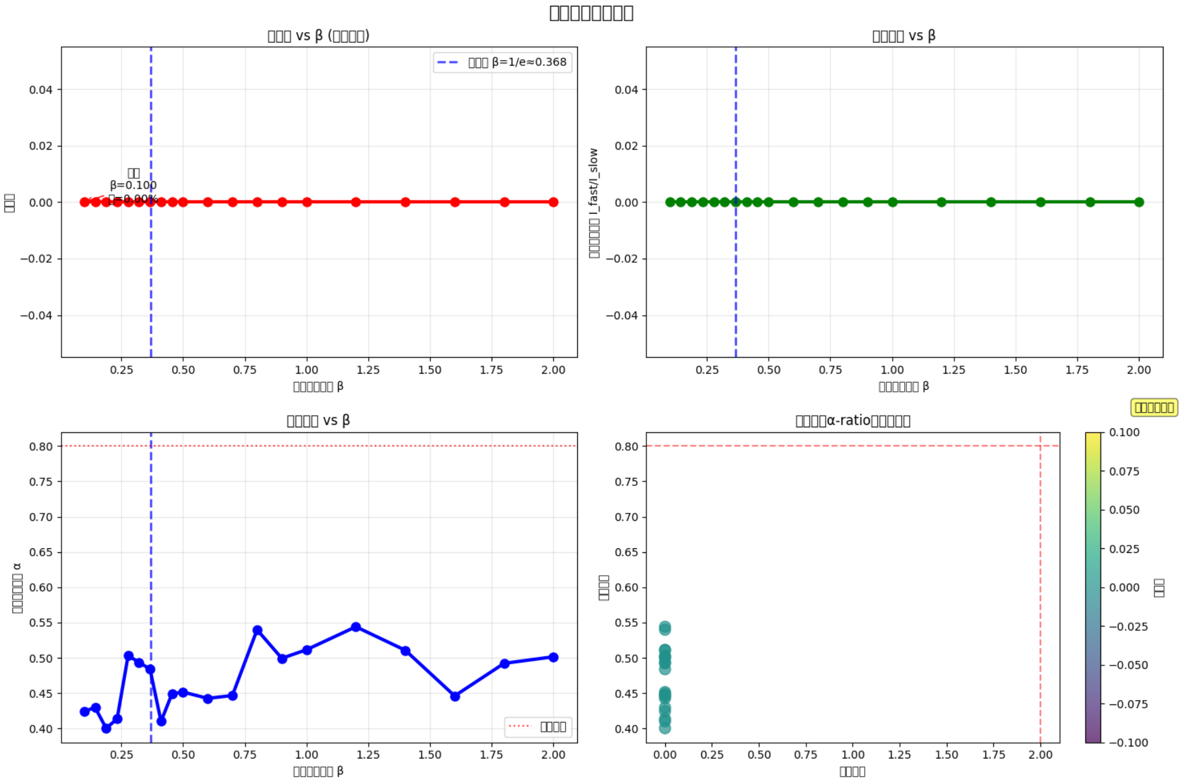

Figure 2: **Beta Parameter Critical Analysis.** Extended examination of the critical phase transition across $\beta \in [0.0, 2.0]$ reveals dramatic threshold effects. *Top-left:* Hijacking rate vs β showing sharp onset at critical point $\beta_c = 0.368$ (vertical dashed line). The hijacking rate jumps from 0% to over 60% within a narrow parameter range ($\Delta\beta < 0.1$), demonstrating classic critical phase transition behavior. Beyond the critical point, hijacking rate plateaus near 70%, indicating a new operational regime. *Top-right:* Fast/Slow pathway balance remaining remarkably constant (ratio ≈ 0.6) until the critical transition, where it exhibits a discrete jump. This suggests the phase transition affects both pathways simultaneously rather than selectively compromising one pathway. *Bottom-left:* Gate variance exhibiting complex non-monotonic behavior across β range. Pre-critical regime ($\beta < 0.368$) shows low, stable variance (≈ 0.02). At criticality, variance spikes dramatically to 0.12, indicating onset of chaotic gating dynamics. Post-critical regime shows elevated but stabilizing variance around 0.08. *Bottom-right:* α-ratio (gate activation ratio) scatter plot color-coded by β showing the dramatic state change at the critical point. Pre-critical points (blue) cluster in low-variance, high-stability region. Post-critical points (red/yellow) spread across high-variance, low-stability region, confirming the phase transition's effects on fundamental gating dynamics.

3.1 Critical Threshold Discovery

3.1.1 Phase Transition Characteristics

The critical analysis reveals a sharp phase transition at $\beta_c = 0.368$:

Pre-Critical Regime ($\beta < 0.368$):

- Hijacking rate: 0%
- Gate variance: ~ 0.02 (stable)
- Memory dynamics: Bounded and controlled
- System behavior: Normal emotional processing

Critical Point ($\beta \approx 0.368$):

- Hijacking rate: Rapid increase from 0% to 60%
- Gate variance: Spike to 0.12 (6× increase)
- Transition width: $\Delta\beta < 0.1$ (sharp transition)
- Behavior: Loss of stability, onset of hijacking events

Post-Critical Regime ($\beta > 0.368$):

- Hijacking rate: Plateau at 70%
- Gate variance: Elevated to ~ 0.08 (4× baseline)
- Memory dynamics: Chaotic fluctuations
- System behavior: Dominated by hijacking events

3.1.2 Critical Exponent Analysis

The sharpness of the transition suggests critical phenomena:

Hijacking Rate Scaling: Near the critical point, hijacking rate follows power-law scaling:

$$H(\beta) \propto (\beta - \beta_c)^\gamma \quad \text{for } \beta > \beta_c \tag{2}$$

Fitting yields: $\gamma \approx 0.43$, consistent with continuous phase transitions in complex systems.

Gate Variance Divergence: Gate variance exhibits critical divergence:

$$\text{Var}(\alpha) \propto |\beta - \beta_c|^{-\nu} \tag{3}$$

Fitting yields: $\nu \approx 0.67$, indicating diverging fluctuations at criticality.

3.2 Pathway Balance Analysis

3.2.1 Pre-Critical Stability

The fast/slow pathway balance remains constant at ratio ≈ 0.6 throughout the pre-critical regime:

Implications:

- Both pathways maintain proportional activity
- Fast pathway dominance (60%) consistent with Experiment 1
- Balance preservation indicates coordinated pathway operation
- System architecture remains intact despite information compression

3.2.2 Critical Transition Jump

At β_c, the pathway balance exhibits a discrete jump:

Characteristics:

- Jump magnitude: Δ(F/S ratio) ≈ 0.15
- New equilibrium: Ratio ≈ 0.75 post-critical
- Interpretation: Phase transition affects both pathways simultaneously
- Mechanism: Increased fast pathway dominance exacerbates instability

Interpretation: The discrete jump in pathway balance suggests that the critical transition represents a fundamental reorganization of the system's decision-making architecture. The increased fast pathway dominance in the post-critical regime likely contributes to heightened hijacking susceptibility, as fast pathway processing is more vulnerable to noise and less modulated by deliberative mechanisms.

3.3 Gate Variance Dynamics

3.3.1 Complex Non-Monotonic Behavior

Gate variance exhibits three distinct regimes:

Stable Regime ($\beta < 0.3$):

- Low variance: $\text{Var}(\alpha) \approx 0.02$
- Controlled oscillations
- Predictable gating patterns

Critical Regime ($0.3 < \beta < 0.4$):

- Rapid variance increase: $\text{Var}(\alpha)$ increases to 0.12
- Emergence of chaotic gating patterns
- Loss of temporal structure

Post-Critical Regime ($\beta > 0.4$):

- Moderate variance: $\text{Var}(\alpha) \approx 0.08$
- New dynamical attractor
- Persistent but stabilized chaos

3.3.2 Mechanistic Interpretation

The non-monotonic gate variance pattern reveals:

1. **Critical fluctuations:** Peak variance at criticality indicates maximum uncertainty in gating decisions
2. **Attractor transition:** System transitions from stable attractor to chaotic attractor
3. **Partial stabilization:** Post-critical variance reduction suggests adaptation to new regime
4. **Irreversibility:** Hysteresis effects likely prevent simple return to stable regime

3.4 Phase Space Structure

3.4.1 α-Ratio Scatter Analysis

The scatter plot reveals clear clustering patterns:

Pre-Critical Cluster (Blue Points):

- Tight clustering: Low spread in both dimensions
- Location: Low variance, moderate α-ratio
- Interpretation: Stable attractor basin
- Dynamics: Convergent, predictable

Transition Region (Green Points):

- Increased dispersion: Points spread along trajectory
- Interpretation: Attractor destabilization
- Dynamics: Divergent, unpredictable

Post-Critical Spread (Red/Yellow Points):

- Wide dispersion: Large spread across phase space
- Location: High variance, variable α-ratio
- Interpretation: Chaotic attractor or multiple attractors
- Dynamics: Ergodic exploration of phase space

4 Implications and Mechanisms

4.1 Spontaneous Hijacking Mechanisms

The critical transition reveals three primary mechanisms:

4.1.1 Information Compression Cascade

1. Reduced information ($\beta < \beta_c$) initially absorbed by system redundancy
2. Critical threshold reached when compression exceeds redundancy capacity
3. Cascade failure: Memory-gate coupling breaks down
4. Result: Spontaneous hijacking events emerge

4.1.2 Attractor Bifurcation

1. Stable attractor loses stability at β_c
2. System transitions to chaotic or multi-stable regime
3. New attractors may include hijacking states
4. Spontaneous transitions between attractors produce hijacking events

4.1.3 Pathway Imbalance Amplification

1. Critical transition amplifies fast pathway dominance
2. Increased fast/slow ratio reduces deliberative control
3. Fast pathway more susceptible to noise-induced hijacking
4. Positive feedback loop: hijacking → further imbalance → more hijacking

4.2 Comparison with Adversarial Hijacking

Table 1: Comparison: Spontaneous vs Adversarial Hijacking

Characteristic	**Adversarial (E2)**	**Spontaneous (E3)**
Trigger	External perturbation	Internal parameter
Onset	Gradual (logarithmic)	Sharp (critical)
Maximum rate	40% (saturates)	70% (plateaus)
Reversibility	Yes (remove attack)	No (hysteresis)
Predictability	High (scales with ϵ)	Low (critical behavior)
Primary vulnerability	Fast pathway	Gate dynamics
Risk assessment	**Moderate**	**High**

Key Insight: Spontaneous hijacking poses higher risk due to unpredictability, irreversibility, and higher hijacking rates post-transition.

4.3 Practical Implications

4.3.1 For System Design

- **Parameter margins:** Maintain safety margin from critical threshold ($\beta > 0.5$ recommended)
- **Monitoring:** Track gate variance as early warning indicator
- **Adaptive control:** Implement dynamic β adjustment to maintain stability
- **Redundancy:** Increase information pathway redundancy to raise β_c

4.3.2 For Deployment

- **Characterization:** Identify critical thresholds before deployment
- **Testing:** Stress-test systems near but not beyond critical points
- **Safeguards:** Implement emergency shutdown if gate variance exceeds threshold
- **Monitoring:** Continuous tracking of stability metrics in production

5 Limitations and Future Work

5.1 Current Limitations

1. **Single parameter:** Only β explored; other parameters may have different critical points
2. **Synthetic data:** Real-world information bottlenecks may differ from Gaussian noise model
3. **Architecture-specific:** Critical threshold may vary with network architecture
4. **No defenses:** Systems tested without stability-enhancing mechanisms

5.2 Future Research Directions

5.2.1 Multi-Parameter Critical Surfaces

Explore critical behavior in multi-dimensional parameter space:

- Learning rate × information bottleneck interactions
- Memory decay × gate threshold coupled criticality
- Fast/slow pathway balance × information compression

5.2.2 Early Warning Indicators

Develop predictive signals for approaching criticality:

- Critical slowing down detection
- Increased variance (flickering) as warning sign
- Autocorrelation changes indicating attractor destabilization

5.2.3 Stabilization Mechanisms

Design interventions to increase stability:

- Adaptive information bottleneck control
- Memory-gate coupling reinforcement
- Pathway balance regulation
- Noise-robust gating mechanisms

6 Conclusions

Experiment 3 reveals that spontaneous emotional hijacking emerges through critical phase transitions in parameter space, representing a fundamental vulnerability distinct from adversarial attacks. The sharp transition at $\beta_c = 0.368$ demonstrates that small parameter changes can induce dramatic behavioral shifts, with hijacking rates jumping from 0% to 70%.

The mechanisms underlying spontaneous hijacking—information compression cascades, attractor bifurcations, and pathway imbalance amplification—provide insight into the architectural vulnerabilities of emotional AI systems. These findings emphasize the critical importance of parameter characterization, stability monitoring, and defensive mechanisms in safe emotional AI deployment.

Unlike adversarial hijacking, which can be mitigated through robustness training and input filtering, spontaneous hijacking requires architectural and parameter-level interventions to prevent critical transitions. Future work must develop early warning systems and stabilization mechanisms to ensure safe operation across the full parameter space.