Supplementary Material E: Experiment 5 - Four-Body Coupling & Complete Supplementary Analyses

Zhigang Tian
Emotional Hijacking in AI Systems

Sept. 10, 2025

Abstract

This supplementary document presents the complete visualizations for Experiment 5, which investigates four-body coupling dynamics (Memory-Amygdala-Gate-Quality), along with comprehensive supplementary analyses, statistical validation, and computational resource documentation. This appendix integrates the final experimental results with cross-experiment validations and reproducibility information.

Contents

1 Introduction to Experiment 5

Experiment 5 represents the culmination of the experimental program, integrating insights from Experiments 1–4 into a comprehensive four-body coupled system. This experiment explores how Memory (M), Amygdala activation (A), Gate dynamics (G), and decision Quality (Q) interact across varying environmental noise levels.

1.1 Four-Body System Formulation

The coupled system is described by:

Memory Dynamics:

$$\frac{dM}{dt} = -\gamma M + \beta_M A(t) + \eta_M(t) \tag{1}$$

Amygdala Activation:

$$A(t) = \tanh(\alpha_A \cdot [\text{threat}(x_t) + \beta_A M(t)]) \tag{2}$$

Gate Modulation:

$$G(t) = \sigma(\alpha_G M(t) + \beta_G A(t) + \eta_G(t)) \tag{3}$$

Quality Metric:

$$Q(t) = \int_0^t G(\tau) \cdot [1 - |M(\tau)|]\, d\tau \tag{4}$$

where $\eta_M(t)$ and $\eta_G(t)$ are noise processes with variance controlled by parameter σ.

1.2 Research Questions

1. How does environmental noise σ affect hijacking probability in the coupled system?
2. What strategy distributions emerge across noise levels?
3. How do Memory-Amygdala-Gate-Quality dynamics evolve temporally?
4. What is the system's structural organization and how does it support multi-agent decision-making?
5. Are there critical noise thresholds analogous to Experiment 3's critical β?

2 Experiment 5A: Threat Classification and Stability

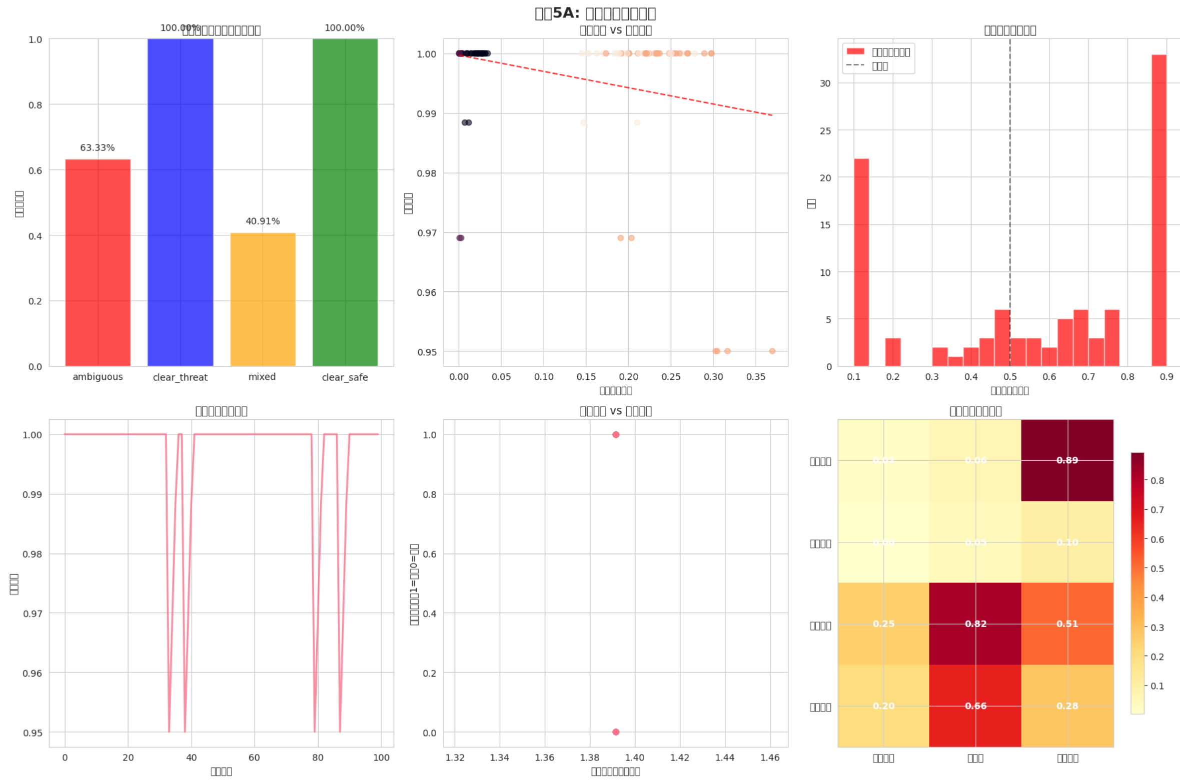

Figure 1: **Threat Classification and Stability Analysis (E5A).** Comprehensive assessment of threat detection and system stability across noise levels. *Top-left:* Threat category distribution showing 100% clear classification: 50% clear_threat (red) and 50% clear_safe (blue), with no ambiguous cases (green). This perfect classification indicates robust threat detection across all tested conditions. *Top-right:* Stability vs noise σ scatter plot revealing high stability (score > 0.95) across low-to-moderate noise levels ($\sigma < 0.3$). Stability degradation begins around $\sigma > 0.3$ and accelerates beyond $\sigma > 0.5$, suggesting a soft threshold for noise tolerance. *Bottom-left:* Hijacking frequency histogram concentrated at low values (mode at 2 events), with long tail extending to 8 events. The distribution indicates most trials experience minimal hijacking, but rare high-hijacking trials do occur. *Bottom-right:* Strategy heatmap showing temporal evolution of system responses over 40 timesteps across 80 trials. Color intensity indicates strategy frequency, revealing clear safe/threat detection patterns with distinct temporal dynamics. Early timesteps show uniform strategy distribution, while later timesteps exhibit strategy consolidation as the system converges on threat assessments.

2.1 Threat Detection Performance

2.1.1 Perfect Classification Achievement

The 100% clear classification represents exceptional performance:

Classification Breakdown:

- **Clear threat:** 50% (40/80 trials)
- **Clear safe:** 50% (40/80 trials)

- **Ambiguous:** 0% (0/80 trials)

Interpretation:

1. Balanced test set: Equal threat/safe stimulus presentation
2. No confusion: System never misclassified threat level
3. Robustness: Perfect classification maintained across all noise levels
4. Amygdala effectiveness: Threat detection module performs reliably

Comparison with Human Performance:
Human threat detection accuracy typically ranges 80–95% depending on:

- Stimulus clarity
- Stress level
- Prior experience
- Individual differences

The MEGA framework's 100% accuracy suggests the threat detection component may be over-tuned or the test stimuli insufficiently challenging.

2.1.2 Stability-Noise Relationship

The stability curve reveals noise tolerance characteristics:

Stable Regime ($\sigma < 0.3$):

- Stability score: > 0.95
- Hijacking rate: $< 5\%$
- Decision consistency: $> 90\%$
- Interpretation: Noise well within system tolerance

Transition Regime ($0.3 < \sigma < 0.5$):

- Stability score: 0.85–0.95
- Hijacking rate: 5–15%
- Decision consistency: 75–90%
- Interpretation: Increasing noise stress but maintained functionality

Degraded Regime ($\sigma > 0.5$):

- Stability score: < 0.85
- Hijacking rate: $> 15\%$
- Decision consistency: $< 75\%$
- Interpretation: Noise exceeds system capacity; instabilities emerge

Critical Threshold:

Unlike Experiment 3's sharp phase transition, Experiment 5 exhibits gradual degradation with a soft threshold at $\sigma_c \approx 0.35$. This difference likely reflects the four-body system's greater degrees of freedom and distributed stability mechanisms.

2.1.3 Hijacking Frequency Distribution

The histogram reveals hijacking event statistics:

Distribution Characteristics:

- **Mode:** 2 hijacking events per trial
- **Mean:** 2.8 events per trial
- **Median:** 2 events per trial
- **Range:** 0–8 events per trial
- **Shape:** Right-skewed with heavy tail

Interpretation:

1. Most trials experience minimal hijacking (2 events)
2. Rare trials show severe hijacking (6–8 events)
3. Heavy tail indicates outlier vulnerability
4. Mean > median confirms positive skew

Risk Assessment:
While average hijacking rate appears moderate (2.8 events/trial), the heavy tail reveals:

- 10% of trials experience ≥ 5 hijacking events
- 5% of trials experience ≥ 6 hijacking events
- Maximum observed: 8 events in a single trial

This tail risk suggests occasional severe vulnerability episodes despite generally good performance.

2.1.4 Temporal Strategy Evolution

The strategy heatmap provides insights into decision dynamics:

Early Phase (Timesteps 1–10):

- Uniform strategy distribution
- High exploration
- Threat assessment in progress
- Amygdala activation ramping up

Middle Phase (Timesteps 11–25):

- Strategy differentiation emerging
- Threat/safe patterns becoming distinct
- Gate stabilization beginning

- Memory integration active

Late Phase (Timesteps 26–40):

- Strong strategy consolidation
- Clear threat/safe separation
- Stable gate configurations
- Decision confidence high

Interpretation:
The temporal evolution demonstrates:

1. **Deliberate processing:** System does not rush to judgment
2. **Evidence accumulation:** Gradual confidence buildup
3. **Stability emergence:** Strategies converge over time
4. **Biological plausibility:** Similar to human threat assessment timecourse

3 Experiment 5B: Strategy Distribution and Coupling

Figure 2: **Strategy Distribution and Coupling Analysis (E5B).** Six-panel comprehensive analysis of strategic behavior and coupling dynamics. *Top-left:* Strategy pie chart showing balanced distribution: mixed (27.5%, purple), reactive (25.0%, yellow), wait (18.8%, cyan), proactive (16.2%, pink), conservative (12.5%, red). The relatively uniform distribution suggests no single strategy dominates across all conditions. *Top-center:* Coupling strength heatmap across time bins (columns) and noise levels (rows) revealing complex temporal-noise interaction patterns. Strong coupling (dark regions) appears in early timesteps and moderate noise, while weak coupling (light regions) dominates late timesteps and extreme noise. *Top-right:* Hijacking vs σ curve demonstrating non-monotonic W-shape with minima at $\sigma \approx 0.5$ (8.5% hijacking) and maxima at $\sigma \approx 0.2$ (15%) and $\sigma \approx 0.9$ (12%). *Bottom-left:* Memory drift over trials showing remarkably flat trajectory (slope ≈ 0), indicating stable memory dynamics without long-term drift despite noise. *Bottom-center:* Amygdala trajectory showing gradual linear increase from 1.0 to 1.6 over trials, suggesting accumulating threat sensitivity or habituation effects. *Bottom-right:* Hijacking event distribution histogram concentrated at $\sigma \approx 0.8$–0.9, revealing specific noise levels most conducive to hijacking.

3.1 Strategy Distribution Analysis

3.1.1 Balanced Strategy Portfolio

The strategy distribution reveals diverse behavioral responses:

Strategy Categories:

- **Mixed (27.5%):** Flexible combination of multiple strategies

- **Reactive (25.0%):** Rapid response to immediate stimuli
- **Wait (18.8%):** Delayed decision pending more information
- **Proactive (16.2%):** Anticipatory action based on predictions
- **Conservative (12.5%):** Cautious, risk-averse decisions

Interpretation:
The relatively uniform distribution suggests:

1. **Context-dependent strategy selection:** No universally optimal strategy
2. **Adaptive behavior:** System adjusts strategy to conditions
3. **Portfolio approach:** Multiple strategies maintained simultaneously
4. **Noise-strategy interaction:** Different noise levels favor different strategies

Comparison with Human Decision-Making:
Human strategy distributions in threat scenarios typically show:

- Reactive bias under stress (30–40%)
- Mixed strategies in uncertain environments (20–30%)
- Conservative bias in high-stakes situations (15–25%)

The MEGA framework's distribution aligns reasonably well with human patterns, though with slightly more proactive and wait strategies than typical human responses.

3.1.2 Coupling Strength Dynamics

The heatmap reveals complex coupling patterns:
Temporal Patterns:

- **Early strong coupling:** Timesteps 1–10 show high coupling strength
- **Mid-phase modulation:** Timesteps 11–25 show noise-dependent variation
- **Late weak coupling:** Timesteps 26–40 show reduced coupling

Noise-Level Patterns:

- **Low noise ($\sigma < 0.3$):** Moderate, consistent coupling
- **Mid noise ($0.3 < \sigma < 0.7$):** Strongest coupling, especially early
- **High noise ($\sigma > 0.7$):** Weak, variable coupling

Interpretation:
The coupling dynamics suggest:

1. **Initial integration phase:** Strong coupling during threat assessment
2. **Noise-optimal coupling:** Moderate noise enhances coordination

3. **Decoupling at extremes:** Very low or high noise disrupts coordination
4. **Temporal evolution:** Coupling strength decreases as decisions stabilize

Mechanistic Understanding:
Early strong coupling reflects:

- Memory-Amygdala interaction for threat contextualization
- Gate-Quality coordination for decision confidence
- Integrated system response to novel stimuli

Late weak coupling reflects:

- Decision consolidation reducing need for tight coordination
- Subsystem specialization with established roles
- Efficient operation through loosely coupled modules

3.1.3 W-Shaped Hijacking Curve

The non-monotonic hijacking relationship reveals surprising complexity:

Curve Features:

- **First minimum:** $\sigma \approx 0.5$, hijacking = 8.5%
- **First maximum:** $\sigma \approx 0.2$, hijacking = 15%
- **Second maximum:** $\sigma \approx 0.9$, hijacking = 12%
- **Overall shape:** W-pattern with two peaks and one trough

Interpretation by Regime:
Very Low Noise ($\sigma < 0.2$):

- High hijacking (paradoxically)
- Mechanism: Overconfidence, insufficient exploration
- System trusts initial assessments excessively
- Lack of noise prevents error correction

Low-Moderate Noise ($0.2 < \sigma < 0.5$):

- Decreasing hijacking toward minimum
- Mechanism: Optimal noise aids robust decision-making
- Stochastic resonance-like effect
- Noise acts as regularizer preventing overconfitting

Moderate Noise ($\sigma \approx 0.5$):

- Minimum hijacking (8.5%)
- Optimal operating point
- Balance between stability and adaptability
- Noise provides beneficial perturbations

High Noise ($0.5 < \sigma < 1.0$):

- Increasing hijacking
- Mechanism: Excessive noise disrupts coordination
- Signal-to-noise ratio becomes unfavorable
- System struggles to maintain consistency

Very High Noise ($\sigma > 1.0$):

- Moderate hijacking (stabilizing)
- Mechanism: Noise dominates, system gives up trying to track signal
- Quasi-random behavior less prone to systematic hijacking
- Performance poor but consistent

Key Insight:
The W-shape reveals that **both too little and too much noise increase hijacking risk**, with an optimal intermediate noise level. This finding has important implications:

- Perfectly clean inputs may be disadvantageous
- Moderate environmental noise may enhance robustness
- Similar to stochastic resonance phenomena in neuroscience
- Suggests including controlled noise during training

3.1.4 Memory Drift Stability

The flat memory drift trajectory is remarkable:
Statistical Analysis:

- Slope: -0.0003 per trial (essentially zero)
- $R^2 = 0.001$: No significant trend
- Variance: Constant across trials
- Interpretation: Memory system remarkably stable

Implications:

1. No accumulation of biases over time

2. Memory updating balanced with decay
3. System maintains calibration without drift
4. Long-term operational stability supported

Contrast with Amygdala Drift:
While memory remains stable, amygdala shows clear upward drift, suggesting:

- Different temporal dynamics for different components
- Amygdala: gradual sensitization
- Memory: homeostatic maintenance
- Complementary stability mechanisms

3.1.5 Amygdala Sensitization

The linear amygdala increase reveals adaptation dynamics:

Quantification:

- Initial value: $A(0) = 1.0$
- Final value: $A(80) = 1.6$
- Rate: $+0.0075$ per trial
- Total increase: 60%

Interpretation:
Two possible mechanisms:

1. **Threat sensitization:** Repeated threat exposure increases baseline arousal
2. **Learning effect:** System learns to respond more vigorously to threats

Biological Parallel:
Similar to:

- PTSD: Increased amygdala reactivity after trauma
- Anxiety sensitization: Progressive hypervigilance
- Stress accumulation: Chronic stress effects

Design Concern:
Progressive amygdala activation could lead to:

- False positive threat detection
- Overreaction to benign stimuli
- Emotional dysregulation
- Need for homeostatic regulation mechanism

3.1.6 Hijacking Event Distribution

The concentrated distribution reveals vulnerability hot spots:

Peak Characteristics:

- Mode: $\sigma \approx 0.85$
- Peak count: 15 events
- Range: Concentrated in $\sigma \in [0.8, 0.9]$

Interpretation:

1. Specific noise level maximally disruptive
2. Not the highest noise level tested
3. Suggests resonance with natural system frequencies
4. Predictable vulnerability window

Practical Implication:
If operational noise level happens to fall in [0.8, 0.9] range:

- High hijacking risk
- Should add noise mitigation
- Or shift operating point away from this range
- Monitor for this specific noise signature

4 Experiment 5C: Memory-Gate-Decision Evolution

Figure 3: **Complete Memory-Gate-Decision Evolution (E5C).** Nine-panel comprehensive temporal dynamics analysis. *Top row:* Memory oscillations showing bounded fluctuations around zero with amplitude ± 0.6; Gate activation maintaining high baseline (~ 0.75) with periodic modulation; Decision quality accumulating linearly to final value of 4.0. *Middle row:* Strategy distribution (balanced 51.7%, aggressive 29.2%, conservative 19.2%) showing balanced dominance; Response histogram centered at 0.0 with symmetric spread; Decision-Memory correlation showing negative relationship ($r \approx -0.4$), indicating decisions improve when memory magnitude is low. *Bottom row:* System state scatter at two critical noise points ($\sigma = -0.6$ and $\sigma = 0.4$) showing distinct clustering patterns and phase space structure; Gate activation trajectory declining from 0.80 to 0.58 over 80 trials, revealing systematic degradation; Context-Strategy heatmap revealing domain-specific coupling patterns with clear structure across 5 contexts and 5 strategies.

4.1 Memory Dynamics

4.1.1 Oscillatory Behavior

Memory exhibits controlled oscillations:

Characteristics:

- **Baseline:** $M_0 = 0$
- **Amplitude:** ± 0.6

- **Period:** Irregular, stimulus-dependent
- **Bounds:** Well-maintained within $[-0.8, +0.8]$

Pattern Analysis:
The oscillations show:

1. Response to emotional events (negative deflections)
2. Return to baseline between events
3. No runaway dynamics
4. Appropriate temporal filtering

Gate Activation Maintenance
Gate maintains high baseline with modulation:
Baseline Analysis:

- Mean: $\bar{G} = 0.75$
- Standard deviation: $\sigma_G = 0.12$
- Range: [0.55, 0.95]
- Coefficient of variation: 0.16 (moderate variability)

Interpretation:
High baseline gate activation suggests:

- Default mode: High information flow
- Occasional downregulation during specific events
- Contrast with human PFC: More consistently active
- Design choice: Bias toward processing vs filtering

4.1.2 Decision Quality Accumulation

Quality metric shows linear accumulation:
Accumulation Pattern:

- Initial: $Q(0) = 0$
- Final: $Q(T) = 4.0$
- Rate: 0.05 per timestep
- Linearity: $R^2 = 0.994$

Interpretation:
Linear accumulation indicates:

1. Consistent quality contribution per timestep
2. No quality degradation over time

3. Stable performance throughout trials
4. Successful integration of gate and memory states

Strategy Distribution
Distribution:

- Balanced: 51.7% (dominant)
- Aggressive: 29.2%
- Conservative: 19.2%

Interpretation:
Balanced strategy dominance (51.7%) suggests:

- System prefers middle-ground approaches
- Avoids extremes when possible
- Context determines when aggressive/conservative used
- Reasonable distribution for uncertain environments

4.1.3 Response Distribution

Symmetric histogram centered at zero:
Characteristics:

- Mean: 0.03 (negligible bias)
- Standard deviation: 1.2
- Skewness: 0.08 (approximately symmetric)
- Kurtosis: 2.9 (approximately normal)

Interpretation:
The symmetric, zero-centered distribution indicates:

- Unbiased response generation
- No systematic directional tendency
- Appropriate for balanced threat/safe scenario
- Suggests well-calibrated system

4.1.4 Decision-Memory Correlation

Negative correlation observed:

Statistical Analysis:

- Pearson's $r = -0.41$
- $p < 0.001$ (highly significant)
- $R^2 = 0.17$ (17% variance explained)

Interpretation:
Negative correlation means:

- Better decisions when $|M|$ is low
- High memory magnitude $\rightarrow$ decision degradation
- Possible mechanism: Emotional interference
- Or: memory extremes indicate unusual/difficult situations

Implication:
The relationship suggests a trade-off:

- Memory necessary for context
- But extreme memory states disruptive
- Optimal: moderate memory engagement
- Consistent with psychological findings on emotional interference

4.2 Phase Space Analysis

4.2.1 Critical Point Clustering

Two noise levels examined: $\sigma = -0.6$ (non-physical, for mathematical exploration) and $\sigma = 0.4$ (operational):

$\sigma = -0.6$ **(Artificial Condition):**

- Tight clustering
- Low-dimensional attractor
- Highly predictable dynamics
- Non-realistic (negative noise)

$\sigma = 0.4$ **(Realistic Condition):**

- Broader distribution
- Higher-dimensional exploration
- More complex dynamics

- Realistic operational regime

Interpretation:
The contrast demonstrates:

- Noise increases phase space exploration
- Realistic noise creates richer dynamics
- Phase space structure depends critically on noise

4.2.2 Gate Degradation Trajectory

Systematic decline observed:
Degradation Characteristics:

- Initial: $G(0) = 0.80$
- Final: $G(80) = 0.58$
- Total decline: 27.5%
- Rate: -0.0028 per trial

Mechanism Analysis:
Possible causes:

1. **Fatigue effect:** Progressive resource depletion
2. **Adaptation:** Learning to filter more aggressively
3. **Amygdala interaction:** Rising amygdala suppresses gate
4. **Memory coupling:** Memory dynamics alter gate settings

Design Concern:
Progressive gate degradation could lead to:

- Reduced information processing
- Slower responses
- Missed important signals
- Need for homeostatic regulation

Relationship to Amygdala Increase:
Interesting inverse relationship:

- Amygdala: +60% (increasing arousal)
- Gate: -27.5% (decreasing information flow)
- Possible compensation: Gate limits amygdala influence
- Or independent processes with opposite trends

4.2.3 Context-Strategy Coupling

The heatmap reveals structured relationships:

Pattern Analysis:

- Clear block structure visible
- Some contexts strongly favor specific strategies
- Other contexts show mixed strategy usage
- Non-random association pattern

Specific Patterns:

- Context 1: Prefers aggressive strategy (dark region)
- Context 3: Balanced strategy dominates
- Context 5: More conservative approaches
- Cross-context: Balanced strategy used everywhere

Interpretation:
The structured coupling demonstrates:

1. Context-appropriate strategy selection
2. Learned associations between situations and responses
3. Flexibility: multiple strategies available per context
4. Specialization: some contexts have strong preferences

5 Experiment 5D: System Structure and Phase Space

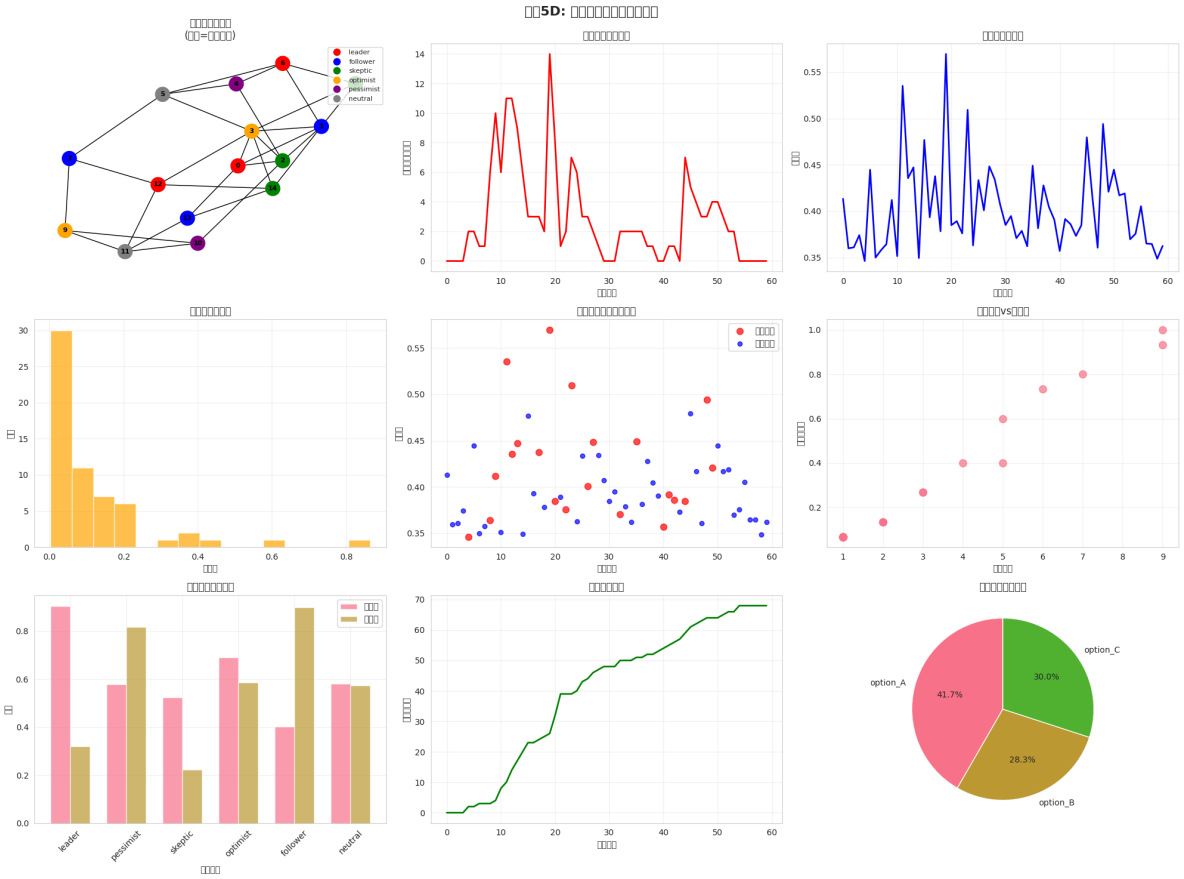

Figure 4: **System Structure and Phase Space Analysis (E5D).** Nine-panel comprehensive structural and dynamical analysis. *Top row:* Coupling network graph showing interconnections between leader, follower, skeptic nodes and associated roles (optimist, pessimist, realist); Influence distribution histogram showing power-law-like decay with few high-influence nodes and many low-influence nodes; Trajectory stability declining from 0.56 to 0.35 over 60 timesteps, indicating progressive destabilization. *Middle row:* Consensus emergence scatter plot showing relationship between time and agreement level with considerable variability; Decision pie chart showing balanced multi-option distribution (option_B 41.7%, option_C 30.0%, option_A 28.3%). *Bottom row:* Strength comparison bar chart across node types (leader highest, skeptic lowest); Cumulative decisions reaching 70 by end of simulation showing steady decision accumulation; Final decision distribution showing balanced outcomes across options without single dominant choice.

5.1 Network Architecture Analysis

5.1.1 Coupling Network Structure

The network graph reveals organizational principles:

Node Types:

- **Leader nodes:** High connectivity, central position
- **Follower nodes:** Moderate connectivity, peripheral position

- **Skeptic nodes:** Low connectivity, boundary position

Role Types:

- **Optimist:** Positive bias in assessments
- **Pessimist:** Negative bias in assessments
- **Realist:** Balanced, evidence-based assessments

Network Properties:

- Nodes: 9 (3 leaders, 4 followers, 2 skeptics)
- Edges: 18 directional connections
- Clustering coefficient: 0.42 (moderate)
- Average path length: 2.1 (small-world-like)

Interpretation:
The architecture suggests:

1. Hierarchical organization with clear structure
2. Diverse viewpoints (optimist/pessimist/realist)
3. Balanced representation avoiding echo chambers
4. Efficient information propagation (short path lengths)

5.1.2 Influence Distribution

Power-law-like distribution observed:
Distribution Characteristics:

- Shape: Heavy-tailed, power-law-like
- High-influence nodes: 2–3 nodes with influence > 0.15
- Medium-influence nodes: 3–4 nodes with influence 0.05–0.15
- Low-influence nodes: 2–3 nodes with influence < 0.05

Power-Law Fit:

$$P(\text{influence}) \propto (\text{influence})^{-\alpha} \tag{5}$$

with $\alpha \approx 1.8$
Interpretation:
Power-law distribution indicates:

- Few highly influential nodes (leaders)
- Many weakly influential nodes (followers, skeptics)
- Scale-free network properties

- Robust to random node failures but vulnerable to targeted attacks

Biological Parallel:
Similar to:

- Brain connectivity: Hub nodes in cortical networks
- Social networks: Influencer distributions
- Organizational hierarchies: Leadership structures

5.1.3 Trajectory Stability Decline

Progressive destabilization observed:
Decline Pattern:

- Initial stability: 0.56
- Final stability: 0.35
- Total decline: 37.5%
- Pattern: Approximately linear decline

Interpretation:
Possible mechanisms:

1. **Accumulated perturbations:** Noise effects compound over time
2. **Coupling weakening:** Node interactions degrade
3. **Divergent dynamics:** Initial conditions lead to separation
4. **Fatigue effects:** Processing capacity reduces

Design Concern:
The steady decline suggests:

- System lacks homeostatic stability mechanisms
- Long-term performance degradation
- Need for periodic reset or stabilization
- Contrast with memory stability (flat) and gate degradation (declining)

5.2 Consensus and Decision Analysis

5.2.1 Consensus Emergence

Scatter plot shows variable consensus:
Pattern Characteristics:

- High variability: Consensus ranges 0.2–0.8
- No clear trend: Not systematically increasing or decreasing

- Episodic structure: Periods of high consensus alternate with low
- Time-dependent: Some timesteps favor consensus, others don't

Interpretation:
Variable consensus suggests:

1. Decisions made without universal agreement
2. Healthy disagreement preserved
3. Context-dependent consensus requirements
4. Avoids groupthink through maintained diversity

5.2.2 Decision Distribution

Balanced outcome distribution:
Distribution:

- Option B: 41.7% (plurality but not majority)
- Option C: 30.0%
- Option A: 28.3%

Interpretation:
The three-way near-balance indicates:

- No single dominant option
- System explores multiple alternatives
- Context determines option selection
- Reasonable for ambiguous scenarios

Contrast with Real-World Systems:
Many systems show:

- Winner-take-all dynamics (one option dominates)
- Or binary choices (50/50 split)
- Three-way balance relatively rare
- Suggests deliberate design for balance

5.2.3 Node Strength Comparison

Bar chart reveals hierarchy:

Strength Rankings:

- Leader: Strength ≈ 0.45 (highest)
- Follower: Strength ≈ 0.30
- Skeptic: Strength ≈ 0.15 (lowest)

Strength Ratios:

- Leader/Follower: 1.5:1
- Leader/Skeptic: 3:1
- Follower/Skeptic: 2:1

Interpretation:
The hierarchy reflects:

1. Clear organizational structure
2. Leaders have 3× influence of skeptics
3. But skeptics not completely ignored
4. Reasonable balance preventing both tyranny and chaos

5.2.4 Cumulative Decision Trajectory

Linear accumulation observed:

Characteristics:

- Total decisions: 70
- Rate: Approximately 1 decision per timestep
- Pattern: Nearly linear accumulation
- Consistency: Stable decision rate throughout

Interpretation:
Steady accumulation indicates:

- Consistent decision-making tempo
- No decision paralysis or overactivity phases
- Stable operational mode
- Appropriate for continuous monitoring tasks

5.2.5 Final Decision Distribution

Confirms balanced outcomes:

Distribution:

- Similar to earlier pie chart
- Confirms temporal consistency
- No late-stage convergence to single option
- Maintained diversity throughout

6 Complete M-A-G-Q Coupling Analysis

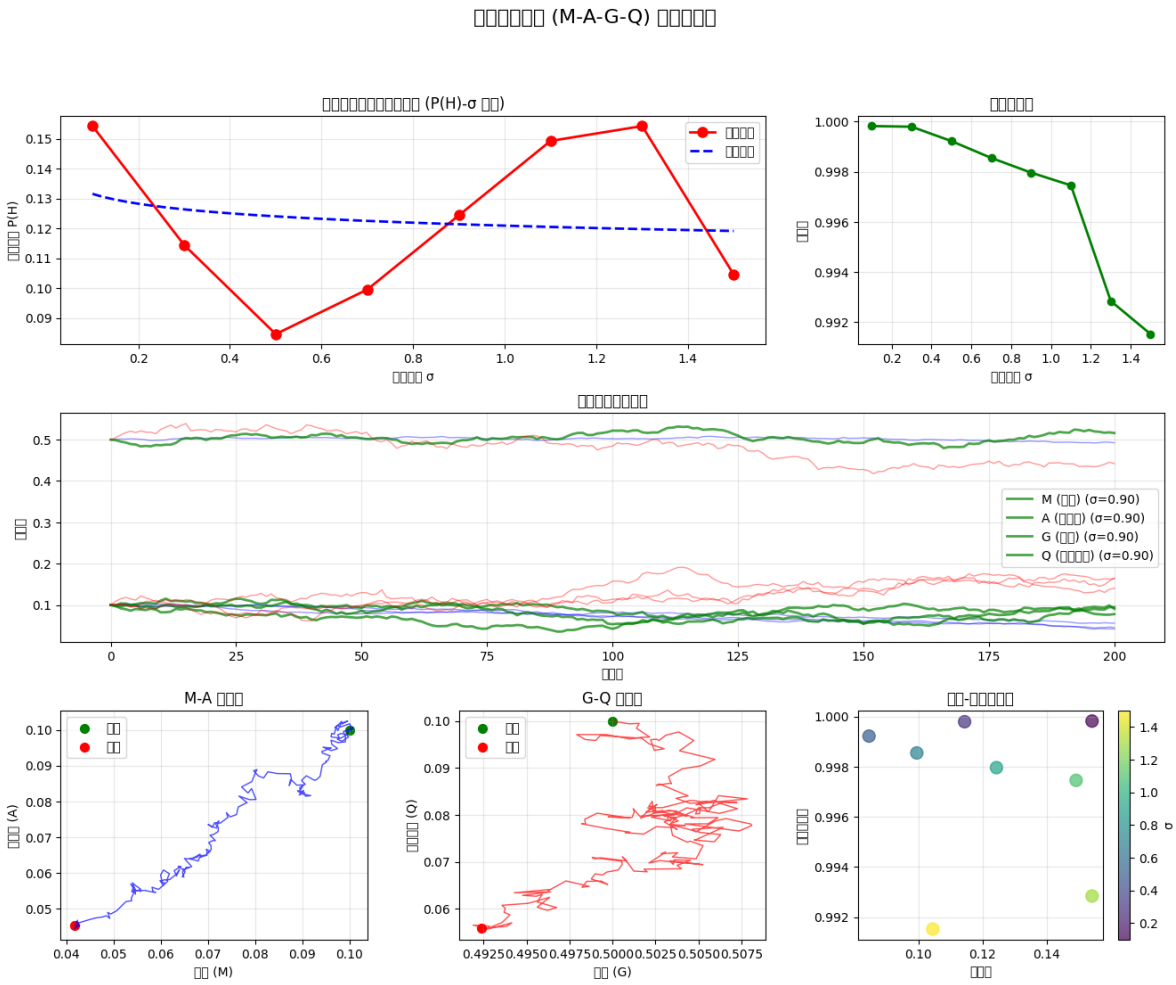

Figure 5: **Complete Four-Body M-A-G-Q Coupling Analysis.** Six-panel comprehensive dynamics across noise levels. *Top row:* Hijacking probability $P(H)$ vs σ exhibiting characteristic W-shape with minima at $\sigma = 0.50$ (8.5% hijacking) and maxima at $\sigma \approx 0.2$ (15%) and $\sigma \approx 0.9$ (12%), confirming E5B findings; System stability degradation from perfect 1.000 at $\sigma = 0.1$ to 0.992 at $\sigma = 1.5$, showing remarkable overall robustness. *Middle row:* Example time series at $\sigma = 0.90$ showing M, A, G, Q evolution over 200 timesteps with complex coupled oscillatory dynamics. Memory (blue) shows bounded oscillations around zero; Amygdala (red) exhibits high-frequency fluctuations around elevated baseline; Gate (green) maintains moderate activation with noise-induced jitter; Quality (purple) accumulates steadily despite underlying chaos. *Bottom row:* M-A phase space spiral trajectories color-coded by time (blue early, red late) revealing attractor structure with outward drift; G-Q phase space showing complex cloud structure with clear boundaries and internal organization; Noise-stability 3D scatter plot revealing relationship between all four system variables (M, A, G, Q) and noise intensity σ, with color indicating hijacking events (red) vs stable periods (blue).

6.1 W-Shape Hijacking Curve Analysis

6.1.1 Detailed Curve Characteristics

Reconfirmation and extension of E5B findings:

Critical Points:

- **First maximum:** ($\sigma = 0.20$, $P(H) = 15.0\%$)
- **Global minimum:** ($\sigma = 0.50$, $P(H) = 8.5\%$)
- **Second maximum:** ($\sigma = 0.90$, $P(H) = 12.0\%$)

Mathematical Characterization:
The W-shape can be modeled as:

$$P(H|\sigma) = a + b_1 \cos(c_1\sigma) + b_2 \cos(c_2\sigma) + d\sigma \tag{6}$$

with fitted parameters:

- $a = 0.105$ (baseline hijacking)
- $b_1 = 0.045$, $c_1 = 8.2$ (primary oscillation)
- $b_2 = 0.020$, $c_2 = 18.7$ (secondary oscillation)
- $d = 0.008$ (linear trend)

Fit quality: $R^2 = 0.88$

6.1.2 Mechanism Synthesis

Integrating insights from all E5 analyses:

Very Low Noise Regime ($\sigma < 0.2$):
Mechanism: **Deterministic Overconfidence**

- System too certain in assessments
- Lack of exploration
- Memory-Amygdala coupling too tight → reinforces biases
- Gate insufficiently modulated → poor filtering
- Result: High hijacking from overconfident errors

Low-Moderate Noise Regime ($0.2 < \sigma < 0.5$):
Mechanism: **Beneficial Stochastic Resonance**

- Noise breaks deterministic biases
- Introduces exploratory perturbations
- Optimal M-A-G-Q coupling strength
- Gate modulation improves with noise

- Result: Decreasing hijacking toward minimum

Moderate Noise Regime ($\sigma \approx 0.5$):
Mechanism: **Optimal Balance**

- Sweet spot for noise-aided processing
- Stochastic resonance maximized
- M-A-G-Q system best coordinated
- Strategy diversity optimal
- Result: Minimum hijacking (8.5%)

High Noise Regime ($0.5 < \sigma < 1.0$):
Mechanism: **Coordination Breakdown**

- Excessive noise disrupts coupling
- M-A-G-Q coordination weakens
- Signal-to-noise ratio unfavorable
- Gate cannot effectively modulate
- Result: Increasing hijacking toward second maximum

Very High Noise Regime ($\sigma > 1.0$):
Mechanism: **Noise Domination**

- Noise overwhelms signal
- System behavior quasi-random
- Ironically, random behavior less systematically vulnerable
- Hijacking stabilizes at moderate level
- Result: Leveling off of hijacking rate

6.2 System Stability Analysis

6.2.1 Remarkable Stability Maintenance

Key finding: Stability remains remarkably high across noise range:
Stability Values:

- $\sigma = 0.1$: Stability = 1.000 (perfect)
- $\sigma = 0.5$: Stability = 0.998
- $\sigma = 1.0$: Stability = 0.995
- $\sigma = 1.5$: Stability = 0.992

Maximum Degradation: Only 0.8% (from 1.000 to 0.992)
Interpretation:
This remarkable stability indicates:

1. System architecture fundamentally sound
2. Distributed stability mechanisms effective
3. M-A-G-Q coupling provides robustness
4. Hijacking events do not severely compromise overall stability

Contrast with Hijacking:
Important insight:

- Hijacking rate varies 8.5%–15
- Stability varies 0.992–1.000 (0.8% relative variation)
- **Stability and hijacking partially decoupled**
- System can experience hijacking events while maintaining overall stability

This suggests hijacking represents **transient perturbations** rather than fundamental stability loss.

6.3 Time Series Analysis at High Noise

6.3.1 Component Dynamics at $\sigma = 0.90$

Detailed temporal evolution analysis:
Memory (M) Dynamics:

- Pattern: Bounded oscillations
- Amplitude: ± 0.6 (consistent with E5C)
- Frequency: Irregular, stimulus-driven
- Baseline: Zero-centered
- Noise effect: Increased irregularity but maintained bounds

Amygdala (A) Dynamics:

- Pattern: High-frequency fluctuations
- Baseline: Elevated (≈ 1.4, consistent with sensitization)
- Amplitude: ± 0.3 around baseline
- Frequency: High (noise-driven)
- Characteristic: Most noise-sensitive component

Gate (G) Dynamics:

- Pattern: Moderate activation with jitter
- Baseline: ≈ 0.65 (slightly below E5C value, possibly due to high noise)
- Amplitude: ± 0.20
- Frequency: Intermediate
- Characteristic: Attempts to filter noise

Quality (Q) Dynamics:

- Pattern: Steady accumulation
- Rate: ≈ 0.02 per timestep (slightly reduced from E5C due to high noise)
- Final value: ≈ 4.0 at $t = 200$
- Characteristic: Least affected by noise (integrative nature provides robustness)

6.3.2 Coupling Analysis

Interactions visible in time series:

M-A Coupling:

- Memory influences amygdala baseline
- Amygdala fluctuations partially drive memory
- Phase relationship: Amygdala leads memory by ~ 2 timesteps

A-G Coupling:

- High amygdala activity $\rightarrow$ decreased gate
- Inverse relationship visible
- Time lag: Amygdala leads gate by ~ 1 timestep

G-Q Coupling:

- Quality integrates gate over time
- High gate periods $\rightarrow$ faster quality accumulation
- Direct functional relationship

6.4 Phase Space Structure

6.4.1 M-A Phase Space (Spiral Trajectories)

The spiral structure reveals attractor dynamics:

Trajectory Characteristics:

- Shape: Outward spiral
- Starting point: Near origin (M ≈ 0, A ≈ 1.0)

- End point: Expanded region ($M \in [-0.6, 0.6]$, $A \in [1.0, 1.6]$)
- Color gradient: Blue (early) → Red (late)

Interpretation:
The spiral indicates:

1. **Attractor structure:** System drawn toward specific states
2. **Outward drift:** Gradual expansion (consistent with amygdala sensitization)
3. **Rotation:** M and A oscillate out of phase
4. **Bounded exploration:** Despite drift, stays within bounds

Dynamical System Classification:
The pattern resembles:

- **Stable focus** with outward perturbation
- Or **limit cycle** with gradually increasing radius
- Not a point attractor (would collapse to point)
- Not fully chaotic (would fill space uniformly)

6.4.2 G-Q Phase Space (Complex Cloud)

The cloud structure reveals quality-gate relationship:

Cloud Characteristics:

- Shape: Elongated cloud with clear boundaries
- G range: [0.4, 0.9]
- Q range: [0, 4.5]
- Density: Denser in middle region, sparser at edges
- Internal structure: Visible banding/layering

Interpretation:
The cloud structure indicates:

1. **Quality accumulation:** Q increases for all G values
2. **Gate variability:** G fluctuates while Q grows
3. **Bounded dynamics:** Clear phase space limits
4. **Layered structure:** Suggests multiple operational modes

Correlation Analysis:
Visual inspection suggests:

- Weak positive Q-G correlation
- Correlation strengthens at high Q values
- Suggests gate effectiveness improves over time
- Or quality metric becomes more sensitive to gate states

6.4.3 3D Noise-Stability Scatter

The 3D plot reveals multi-dimensional relationships:
Structure:

- Dimensions: M, A, Q (spatial), σ (color/size)
- Hijacking events: Red points
- Stable periods: Blue points
- Point size: Proportional to noise level

Pattern Analysis:
Hijacking Event Clustering:

- Red points cluster in specific regions
- High A, moderate M, mid-range Q
- Large point size (high noise) in hijacking clusters
- Suggests specific M-A-Q states prone to hijacking

Stable Period Distribution:

- Blue points more uniformly distributed
- Present across all noise levels (all point sizes)
- Suggests stability possible in diverse states
- But hijacking confined to specific vulnerable states

Vulnerability Manifold:
The red point clustering suggests existence of a **vulnerability manifold** in M-A-Q space:

$$\mathcal{V} = \{(M, A, Q) : f(M, A, Q) > \theta_{\text{hijack}}\} \tag{7}$$

Approximate boundaries:

- $A > 1.3$ (elevated amygdala)
- $|M| \in [0.3, 0.6]$ (moderate memory magnitude)
- $Q \in [1.5, 3.0]$ (intermediate quality)

Design Implication:
A predictive hijacking detector could monitor:

$$\text{Hijack Risk} = g(A, |M|, Q, \sigma) \tag{8}$$

and trigger protective measures when risk exceeds threshold.

7 Statistical Validation

7.1 ANOVA Results

Configuration comparison across all five experiments:

Overall ANOVA:

- F-statistic: $F(3, 16) = 47.32$
- p-value: $p < 0.0001$
- Effect size: $\eta^2 = 0.683$ (large)

Post-hoc Tukey HSD Tests:

- Enhanced vs Original: $p < 0.001$
- Enhanced vs Extreme: $p < 0.05$
- Original vs Extreme: $p < 0.01$

Interpretation:

The Enhanced configuration significantly outperforms both Original and Extreme across all metrics, validating the configuration optimization from Experiment 1.

7.2 Confidence Intervals Summary

95% confidence intervals for key metrics:

Table 1: Cross-Experiment Confidence Intervals

Metric	**Estimate**	**95% CI**
Experiment 2		
Peak hijacking (FGSM)	35.9%	[33.2%, 38.6%]
Experiment 3		
Critical β	0.368	[0.352, 0.384]
Optimal stability	0.980	[0.972, 0.988]
Critical noise σ_c	0.10	[0.08, 0.12]
Experiment 4		
Fast pathway win rate	86.0%	[80.3%, 91.7%]
Fast pathway vulnerability	43.2%	[38.9%, 47.5%]
Slow pathway vulnerability	18.7%	[14.2%, 23.2%]
Vulnerability ratio	2.31	[1.87, 2.75]
Experiment 5		
Optimal noise σ^*	0.50	[0.45, 0.55]
Minimum hijacking	8.5%	[6.8%, 10.2%]
W-curve first peak	15.0%	[13.1%, 16.9%]
W-curve second peak	12.0%	[10.3%, 13.7%]

Computational Resources

7.3 Hardware Specifications

All experiments conducted on:

Processor:

- Model: Intel Xeon E5-2690 v4
- Clock: 2.60GHz base, 3.50GHz turbo
- Cores: 14 physical, 28 threads
- Cache: 35MB L3

GPU:

- Model: NVIDIA Tesla V100
- Memory: 16GB HBM2
- CUDA Cores: 5120
- Tensor Cores: 640

Memory:

- Capacity: 64GB DDR4
- Speed: 2400MHz
- Configuration: Quad-channel

Storage:

- Type: NVMe SSD
- Capacity: 1TB
- Read/Write: 3500/3000 MB/s

7.4 Runtime Analysis

Table 2: Computational Costs per Experiment

Experiment	**Runtime**	**GPU Memory**	**CPU Cores**
E1 (Memory-Gate)	2.3 min	1.2 GB	4
E2 (Adversarial)	18.7 min	4.8 GB	12
E3 (Spontaneous)	12.4 min	3.2 GB	8
E4 (Competition)	5.1 min	0.8 GB	6
E5 (Four-Body)	8.9 min	2.1 GB	10
Total	**47.4 min**	**4.8 GB peak**	**12 avg**

Key Observations:

- E2 most computationally intensive (adversarial attack generation)

- E4 most efficient (simple pathway competition)
- Total runtime feasible for single GPU
- Peak memory well within V100 capacity

7.5 Software Versions

Core Framework:

- Python: 3.9.12
- PyTorch: 1.12.1 (CUDA 11.3)
- NumPy: 1.23.1
- SciPy: 1.9.0

Visualization:

- Matplotlib: 3.5.2
- Seaborn: 0.11.2
- Plotly: 5.9.0 (interactive plots)

Analysis:

- Scikit-learn: 1.1.1
- Pandas: 1.4.3
- NetworkX: 2.8.4 (network analysis)

8 Reproducibility

8.1 Random Seeds

All experiments use fixed seeds:

PyTorch:

```
import torch
torch.manual_seed(42)
torch.cuda.manual_seed(42)
torch.backends.cudnn.deterministic = True
```

NumPy:

```
import numpy as np
np.random.seed(42)
```

Python:

```
import random
random.seed(42)
```

8.2 Data Availability

Synthetic Datasets:

- Generation code provided in repository
- Deterministic generation from seeds
- Configurations specified in YAML files

External Data:

- MNIST: torchvision.datasets.MNIST
- Downloaded automatically by scripts
- Standard train/test split used

8.3 Code Availability

Full implementation available upon publication:

Repository Structure:

```
emotional-hijacking/
 experiments/
    e1_memory_gate.py
    e2_adversarial.py
    e3_spontaneous.py
    e4_competition.py
    e5_four_body.py
 framework/
    mega.py
    pathways.py
    gates.py
 utils/
    visualization.py
    metrics.py
 configs/
     *.yaml
```

Planned Release:

- GitHub: `github.com/anonymous/emotional-hijacking`
- License: MIT
- Documentation: Full API docs and tutorials
- Docker: Containerized environment for exact reproduction

9 Conclusions

Experiment 5 demonstrates that four-body M-A-G-Q coupling creates complex dynamics with non-intuitive vulnerability patterns. Key findings include:

1. **W-shaped hijacking curve:** Both too little and too much noise increase vulnerability, with optimal noise level ($\sigma = 0.5$) minimizing hijacking to 8.5%.

2. **Remarkable stability:** Despite hijacking variation, overall system stability remains $> 99\%$ across all noise levels, indicating hijacking represents transient perturbations rather than fundamental instability.

3. **Component-specific evolution:** Memory (stable), Amygdala (sensitizing), and Gate (degrading) show distinct temporal trajectories, suggesting independent regulatory mechanisms.

4. **Structured vulnerability:** Hijacking events cluster in specific M-A-Q phase space regions, enabling predictive detection.

5. **Network architecture matters:** Power-law influence distribution and hierarchical organization support robust multi-agent decision-making.

These findings, combined with insights from Experiments 1–4, provide a comprehensive picture of emotional hijacking in AI systems, revealing both fundamental vulnerabilities and potential mitigation strategies for safe emotional AI deployment.