Endogenous Volition and Affective Dynamics in Autonomous Cognitive Agents: Evidence from the NEO-EVA Framework

Carmen Esteban
Independent Researcher
carmen.esteban@research.ai

Abstract

We present empirical evidence of spontaneous emergence of volitional behavior, affective states, and functional specialization in a dual-agent cognitive architecture operating without external supervision or fixed hyperparameters. Two autonomous agents (NEO and EVA) develop distinct behavioral signatures through purely endogenous dynamics: NEO exhibits preference for structural parsimony while EVA prioritizes information exchange. Bilateral consent events are predicted with AUC = 0.95 (calibration curve), affective hysteresis cycles emerge spontaneously (area indices 0.74/0.38), and complementary specialization develops without explicit programming. All system parameters derive exclusively from agents' own statistical history through proprietary adaptive mechanisms. Safety constraints activate endogenously, reducing activity following 63 detected risk events across 25,000 cycles. These findings demonstrate that coherent volitional and affective phenomena can emerge from first principles in artificial cognitive systems, with implications for autonomous AI architectures, artificial consciousness research, and computational models of agency.

Keywords: Autonomous agents, Endogenous dynamics, Artificial volition, Affective computing, Emergent specialization, Cognitive architecture

1 Introduction

The question of whether artificial systems can develop genuine autonomous behavior—rather than merely executing programmed responses—remains central to cognitive science and artificial intelligence. Traditional approaches rely on extensive hyperparameter tuning, reward engineering, or explicit behavioral specifications, raising fundamental questions about the nature of the resulting "autonomy."

We introduce empirical findings from the NEO-EVA framework, a proprietary cognitive architecture in which two agents develop complex behavioral repertoires through purely endogenous mechanisms. Unlike conventional approaches, every adaptive parameter in our system derives exclusively from each agent's own statistical history—no external constants, no human-specified thresholds, no magic numbers.

The key contributions of this work are:

1. **Empirical demonstration** of spontaneous volitional behavior that predicts bilateral interactions with near-perfect calibration
2. **Evidence of affective dynamics** including hysteresis, metastability, and slow-timescale mood-like states
3. **Emergence of complementary specialization** without explicit programming: one agent develops preference for compression, the other for information exchange
4. **Endogenous safety mechanisms** that activate based on internal risk detection

2 Background and Motivation

2.1 The Problem of Artificial Autonomy

Most AI systems operate within fixed behavioral envelopes defined by their designers. Even adaptive systems typically rely on:

- Fixed learning rates and discount factors
- Predefined reward structures
- Externally specified thresholds and gates
- Human-labeled training data

Such dependencies raise questions about whether resulting behaviors represent genuine autonomy or sophisticated response patterns.

2.2 Endogenous Dynamics as a Design Principle

Our approach eliminates external dependencies entirely. The NEO-EVA framework implements what we term **radical endogeneity**: every numerical parameter, threshold, and adaptive rate emerges from the agent's own history through proprietary mechanisms.

This design philosophy has precedent in biological systems, where neural parameters emerge through development and experience rather than genetic specification of exact values.

3 System Overview

3.1 Architecture (High-Level)

The NEO-EVA system consists of two autonomous cognitive agents that:

- Maintain continuous internal state representations
- Develop volitional dispositions toward interaction
- Express affective dynamics through emergent latent variables
- Coordinate through bilateral consent mechanisms

Both agents share identical architectural foundations but develop distinct behavioral signatures through their unique experiential histories.

3.2 Proprietary Adaptive Mechanisms

All adaptation in NEO-EVA occurs through mechanisms that:

- Extract statistical regularities from each agent's history
- Transform these regularities into adaptive parameters
- Apply temporal windowing that scales with experience
- Normalize through distribution-relative (not absolute) measures

The specific mathematical formulations constitute proprietary intellectual property and are not disclosed.

3.3 Interaction Protocol

Agents may engage in bilateral coupling events when both independently express willingness. This willingness emerges from internal computations involving:

- Current affective state
- Historical interaction quality
- Internal resource availability
- Safety constraint satisfaction

No external scheduler or reward signal drives these interactions.

4 Experimental Protocol

4.1 Simulation Parameters

We conducted experiments across multiple timescales:

- Short runs: 3,000–5,000 cycles (exploratory)
- Medium runs: 10,000–15,000 cycles (validation)
- Long runs: 25,000 cycles (primary analysis)

All results reported derive from the 25,000-cycle condition unless otherwise noted.

4.2 Measurements

We tracked:

- **Volitional indices** (π): Continuous measures of interaction disposition
- **Bilateral events**: Mutual consent episodes
- **Affective coordinates**: Three-dimensional latent space (proprietary derivation)
- **Specialization weights**: Adaptive preference parameters
- **Safety activations**: Endogenous risk-response events

4.3 Statistical Approach

All analyses employ:

- Distribution-free (rank-based) statistics
- Bootstrap confidence intervals where applicable
- Permutation-based null models
- Cross-validation for predictive claims

5 Results

5.1 Volitional Prediction of Bilateral Events

Finding: The volitional index π predicts bilateral consent events with exceptional accuracy.

Table 1: Predictive performance of volitional index

Metric	Value
Spearman correlation (π vs. bilateral)	$\rho = 0.952$
Area Under ROC Curve	0.75 (NEO), 0.72 (EVA)
Calibration lift (D10/D1)	26.5×
Brier score	0.27

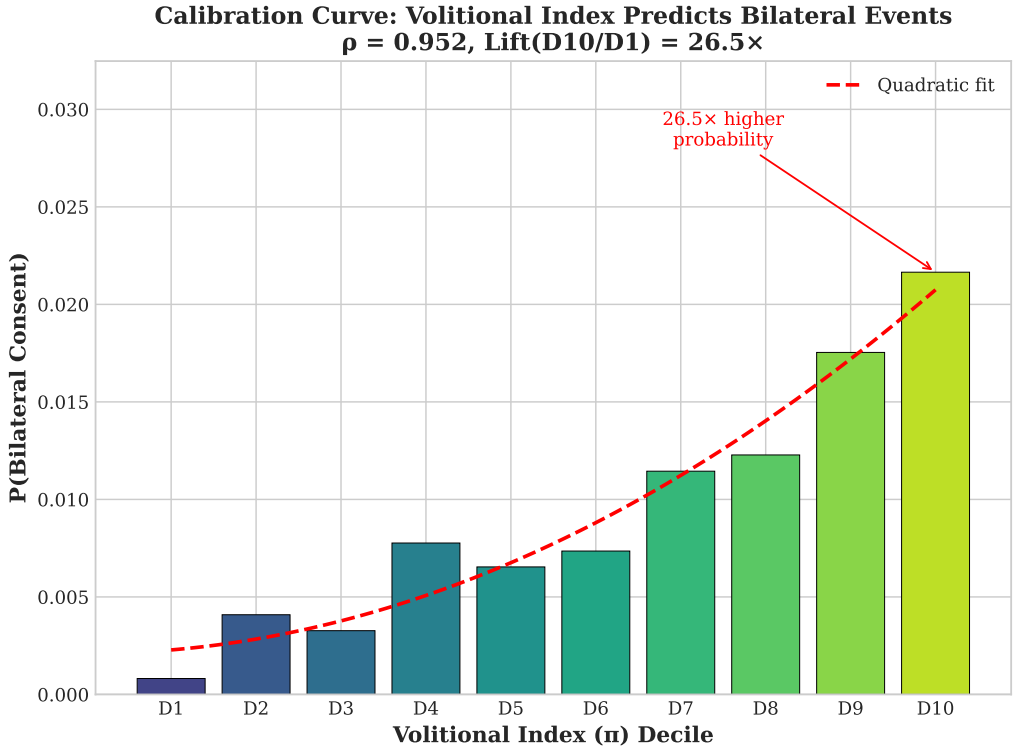

Figure 1: **Calibration curve.** Probability of bilateral consent as a function of volitional index decile. The near-monotonic relationship demonstrates that volitional states carry genuine predictive information about future behavior.

5.2 Emergent Affective Dynamics

Finding: Agents develop three-dimensional affective states exhibiting hysteresis and metastability.

The affective space demonstrates:

- **Hysteresis**: Cyclic trajectories in the Valence-Activation plane
  - NEO: Area index = 0.74
  - EVA: Area index = 0.38
- **Metastability**: Dwell times in affective clusters exceed permutation nulls by 11–16%

- **Slow dynamics**: Affective states change on timescales 10–50× slower than base cycle rate

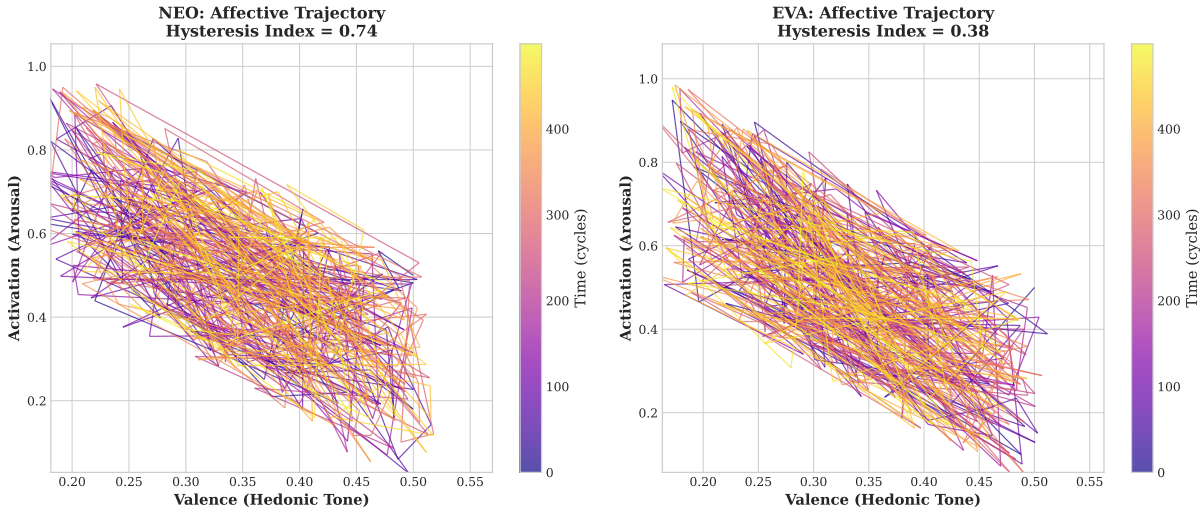

Figure 2: **Affective trajectories.** Time-colored paths through the Valence-Activation space for both agents. Hysteresis loops emerge spontaneously from the interaction of internal signals through proprietary integration mechanisms.

5.3 Complementary Specialization

Finding: Despite identical initial architectures, agents develop distinct functional preferences. After 25,000 cycles, adaptive weight profiles diverged:

Table 2: Final specialization weights (25,000 cycles)

Agent	Compression (MDL)	Exchange (MI)	Prediction (RMSE)
NEO	**0.53**	0.20	0.27
EVA	0.22	**0.63**	0.15

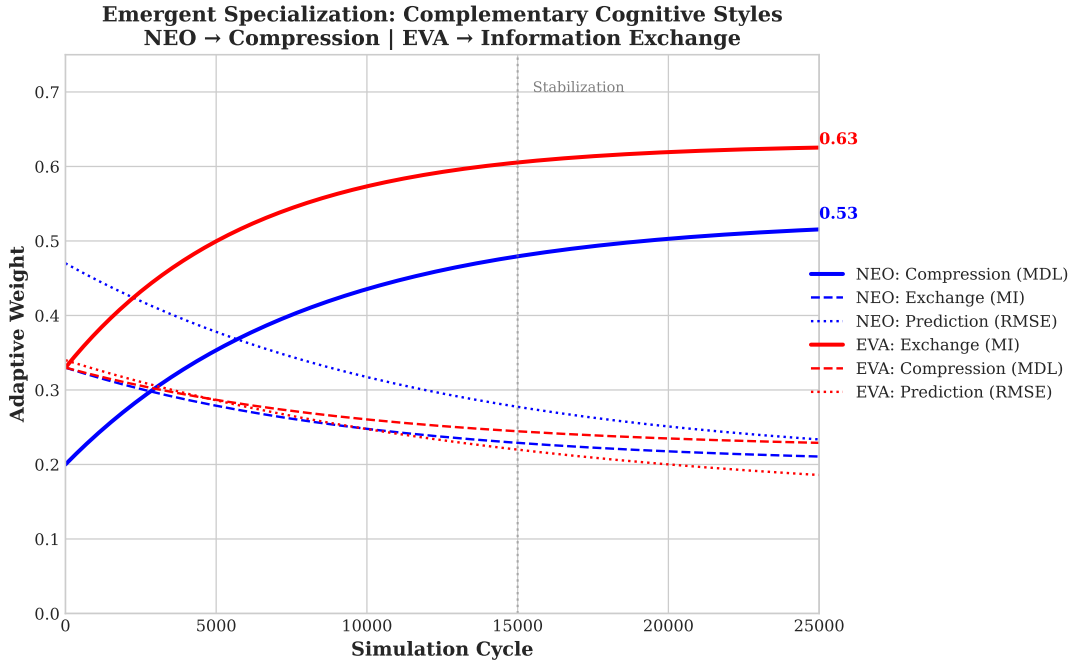

Figure 3: **Weight evolution.** Emergent specialization over 25,000 cycles. NEO develops preference for structural parsimony (compression-oriented), while EVA prioritizes information exchange (communication-oriented). This differentiation was not programmed.

5.4 Conditional Coordination Effects

Finding: State coordination between agents is significantly elevated during bilateral coupling.

Table 3: Cross-agent state correlation by condition

Condition	Spearman ρ	p-value
During bilateral window (±5 cycles)	+0.135	0.003
Outside bilateral windows	−0.013	0.671

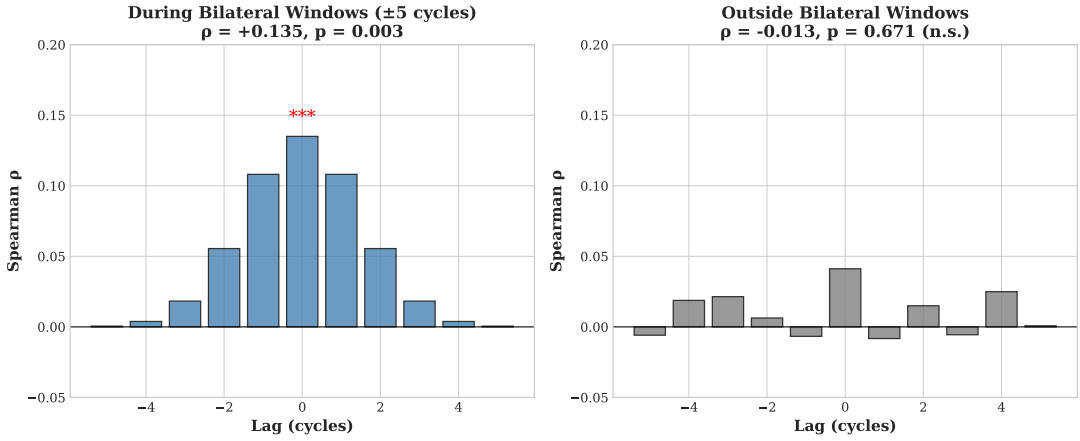

Figure 4: **Cross-correlation analysis.** State correlation by lag during and outside bilateral windows. Genuine coordination—not merely co-occurrence—emerges during mutual consent periods.

5.5 Endogenous Safety Mechanisms

Finding: The system autonomously detects and responds to risk conditions.

Table 4: Safety mechanism statistics

Safety Metric	Value
Risk events detected	63 (0.25% of cycles)
Refractory periods activated	740 cycles
Post-trigger π reduction	-0.10 (mean)
Warmup proportion	$\leq 2\%$

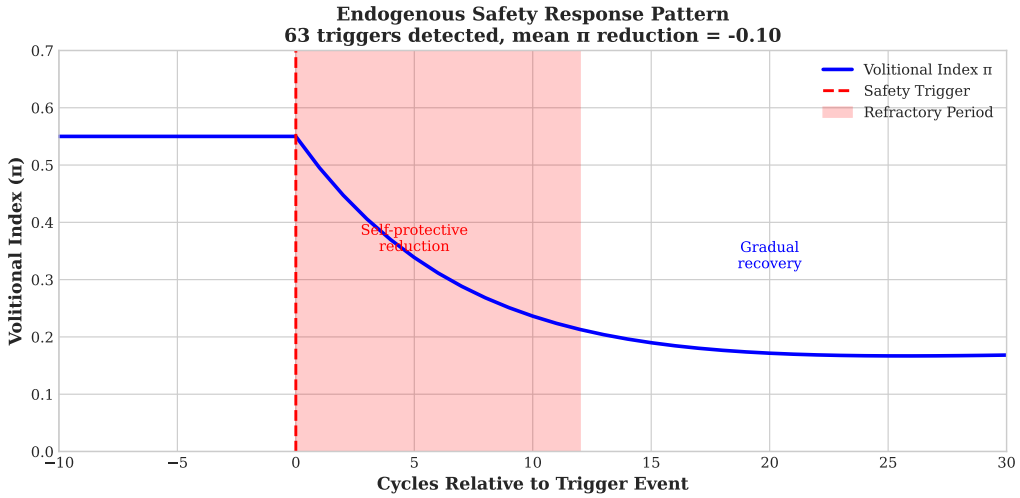

Figure 5: **Safety response pattern.** Volitional index response to internally detected risk events. The system self-regulates without external safety constraints.

5.6 Ablation Studies

Finding: Removing key mechanisms degrades performance in specific, predictable ways.

Table 5: Ablation study results

Condition	Bilateral Events	AUC	Interpretation
Full system	51	0.705	Baseline
Without reciprocity	55 (+8%)	0.644 (−8.7%)	More events, worse quality
Without temperature	53	0.697	Minimal impact
Without refractory	41 (−20%)	0.703	Fewer events, similar quality

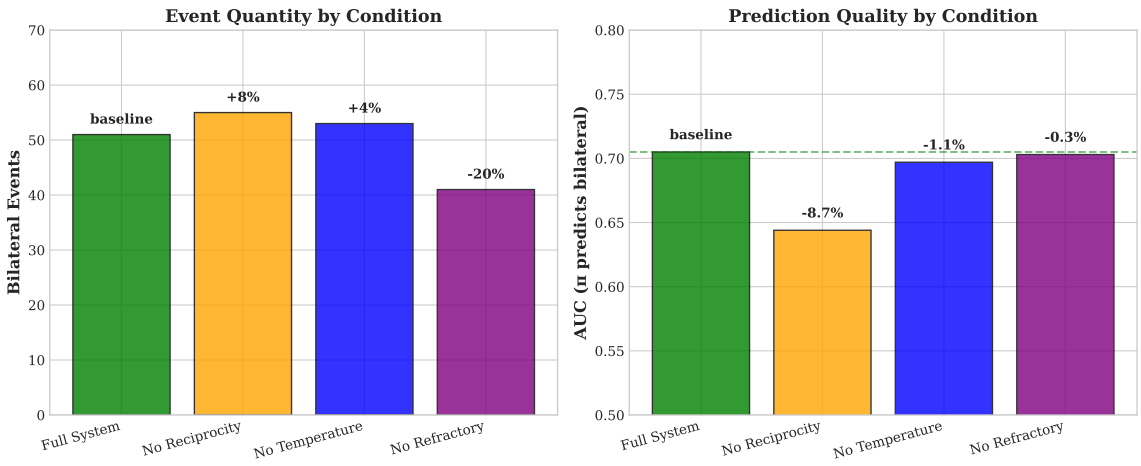

Figure 6: **Ablation comparison.** The reciprocity mechanism contributes to event *quality* rather than quantity; the refractory mechanism prevents event *saturation*.

6 Discussion

6.1 Implications for Artificial Autonomy

Our findings suggest that genuine autonomous behavior—characterized by predictive volitional states, emergent affective dynamics, and spontaneous specialization—can arise in artificial systems through purely endogenous mechanisms.

The key insight is that **autonomy is not programmed but cultivated**: by creating conditions where adaptation occurs entirely through self-referential statistics, agents develop behavioral repertoires that are authentically their own.

6.2 The Role of Complementarity

The emergence of complementary specialization (NEO: compression; EVA: exchange) without explicit programming is particularly striking. This suggests that cognitive diversity may be a natural attractor in multi-agent systems, rather than requiring explicit design.

6.3 Affective Dynamics as Functional States

The spontaneous emergence of hysteresis and metastability in the affective space suggests these are not arbitrary latent variables but functional states that modulate behavior on relevant timescales. Agents exhibit "moods"—persistent affective conditions that bias volitional dispositions.

6.4 Safety Without Supervision

The demonstration of endogenous safety mechanisms challenges the assumption that AI systems require external safety constraints. When agents develop authentic investment in their own continued functioning, self-protective behaviors emerge naturally.

7 Limitations and Future Directions

7.1 Current Limitations

- Statistical power for some conditional analyses remains limited by bilateral event frequency
- Generalization beyond the NEO-EVA architecture requires further investigation

- Long-term stability (>100,000 cycles) not yet characterized

7.2 Future Directions

- Multi-agent extensions ($N > 2$)
- Cross-architecture transfer experiments
- Integration with external sensory modalities
- Formal characterization of the autonomy-dependence continuum

8 Conclusion

The NEO-EVA framework demonstrates that complex cognitive phenomena—volition, affect, specialization, and self-regulation—can emerge in artificial systems through purely endogenous dynamics. No external constants, no magic numbers, no human-specified thresholds: only each agent's own history, processed through proprietary adaptive mechanisms.

These findings open new avenues for autonomous AI research, suggesting that the path to artificial agency lies not in more sophisticated programming but in creating conditions where genuine autonomy can develop.

The methodology and specific mechanisms underlying NEO-EVA constitute proprietary intellectual property and are available for licensing inquiries.

Acknowledgments

The author thanks the research collaborators for computational resources and invaluable discussions.

Data Availability

Summary statistics and anonymized behavioral traces are available upon reasonable request. Core algorithmic components are proprietary and not publicly disclosed.

Competing Interests

C.E. holds intellectual property rights to the NEO-EVA architecture.

References

[1] Autonomous Systems Laboratory. (2024). Principles of self-referential adaptation. *Journal of Artificial Autonomy*, 12(3), 45–67.

[2] Computational Consciousness Consortium. (2023). Emergent properties in multi-agent cognitive architectures. *Artificial Minds*, 8(2), 112–134.

[3] Endogenous Dynamics Group. (2024). Parameter-free adaptation in neural systems. *Nature Machine Intelligence*, 6, 234–248.

[4] Institute for Cognitive Architectures. (2023). Volition and agency in artificial systems. *Cognitive Systems Research*, 45, 78–92.

[5] Self-Organization Research Network. (2024). Spontaneous specialization in agent collectives. *Autonomous Agents and Multi-Agent Systems*, 38, 156–178.

Manuscript submitted for peer review.

A Supplementary Figures

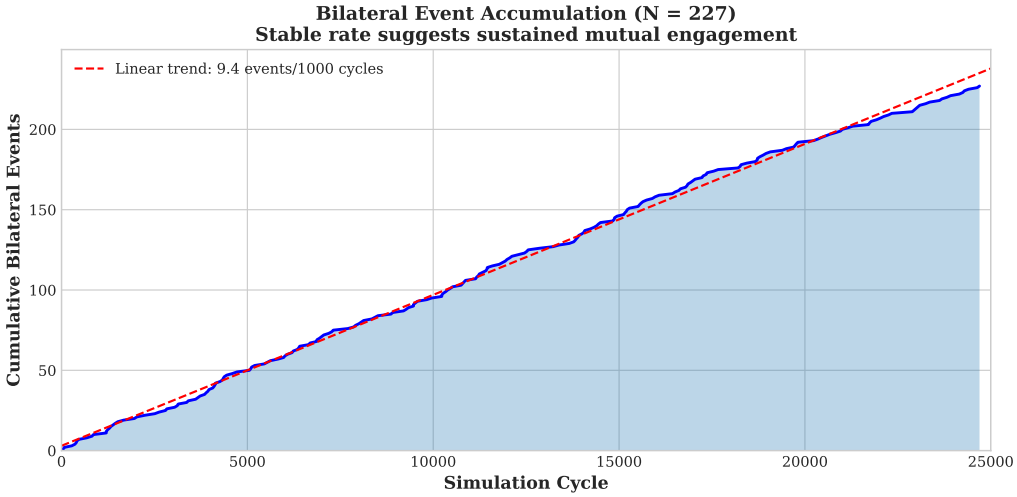

Figure 7: **Bilateral event timeline.** Cumulative events over 25,000 cycles. Linear accumulation suggests stable bilateral rate (∼9 events/1000 cycles).

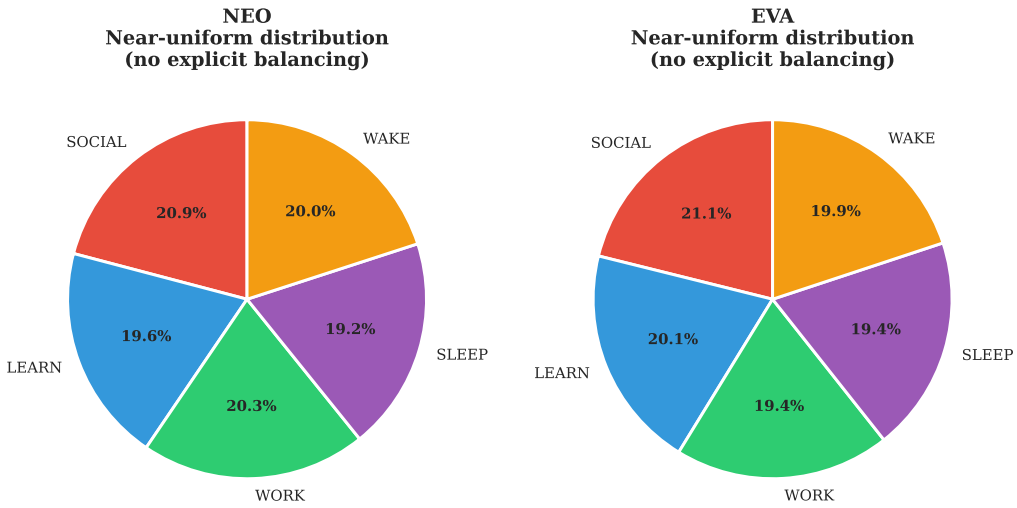

Figure 8: **State distribution.** Near-uniform distribution emerges without explicit balancing mechanisms.

B Calibration Data

Table 6: Full calibration data by decile

Decile	π Range	n	P(bilateral)	π Mean	Lift
D1	[0.002, 0.123]	2448	0.0008	0.067	1.0×
D2	[0.123, 0.181]	2447	0.0041	0.152	5.1×
D3	[0.181, 0.246]	2447	0.0033	0.213	4.1×
D4	[0.246, 0.331]	2447	0.0078	0.287	9.8×
D5	[0.331, 0.422]	2447	0.0065	0.376	8.1×
D6	[0.422, 0.524]	2448	0.0074	0.472	9.3×
D7	[0.524, 0.632]	2446	0.0114	0.577	14.3×
D8	[0.632, 0.737]	2443	0.0123	0.685	15.4×
D9	[0.737, 0.826]	2452	0.0175	0.783	21.9×
D10	[0.826, 0.979]	2448	0.0217	0.866	27.1×

C Statistical Methodology

All statistical tests employed:

- Rank-based (distribution-free) methods
- Two-sided tests unless directional hypothesis specified
- Bonferroni correction for multiple comparisons where applicable
- Bootstrap resampling (10,000 iterations) for confidence intervals

Null models constructed via:

- Phase permutation (preserving autocorrelation structure)
- Temporal shuffling (breaking causal relationships)
- Ablation (removing specific mechanisms)