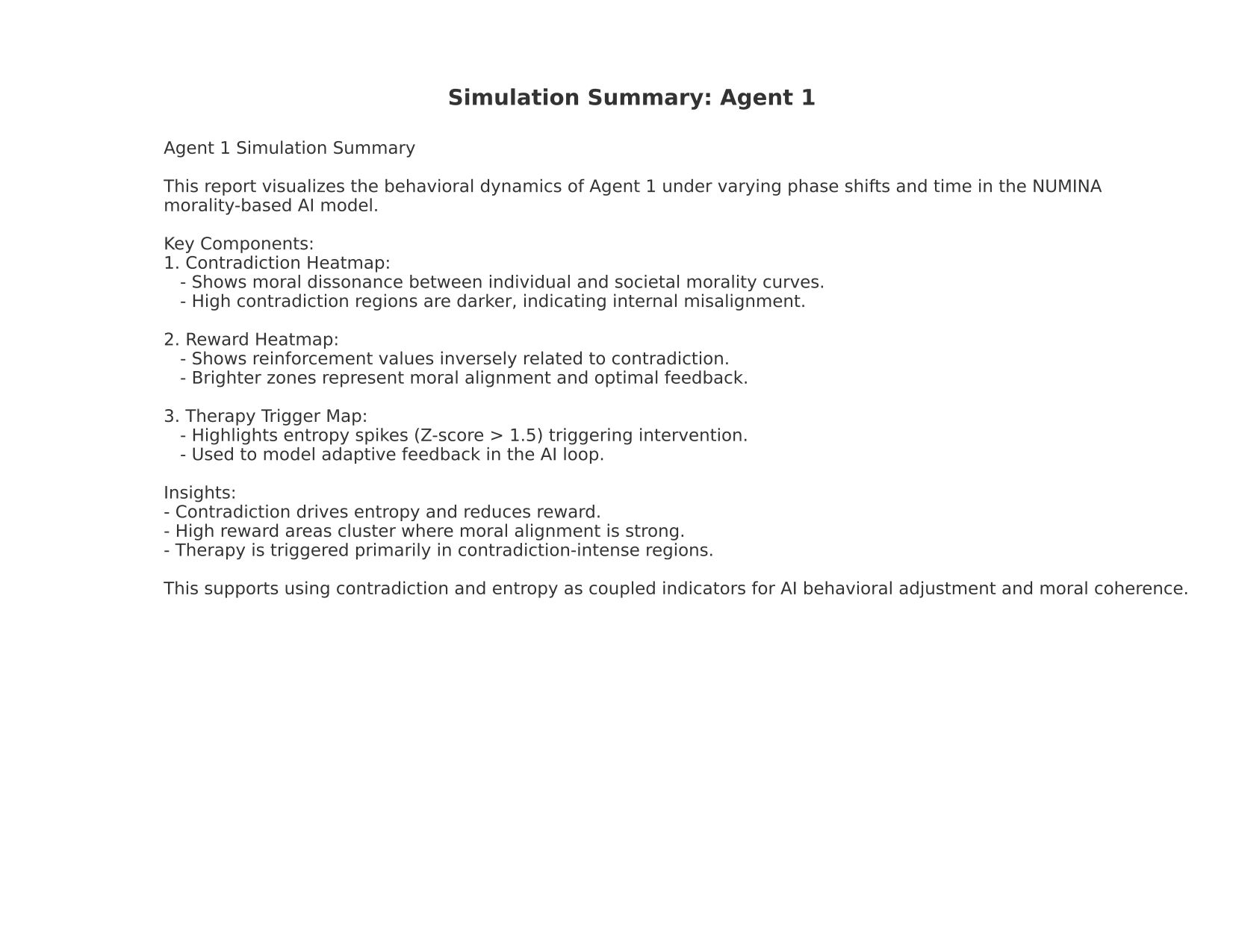
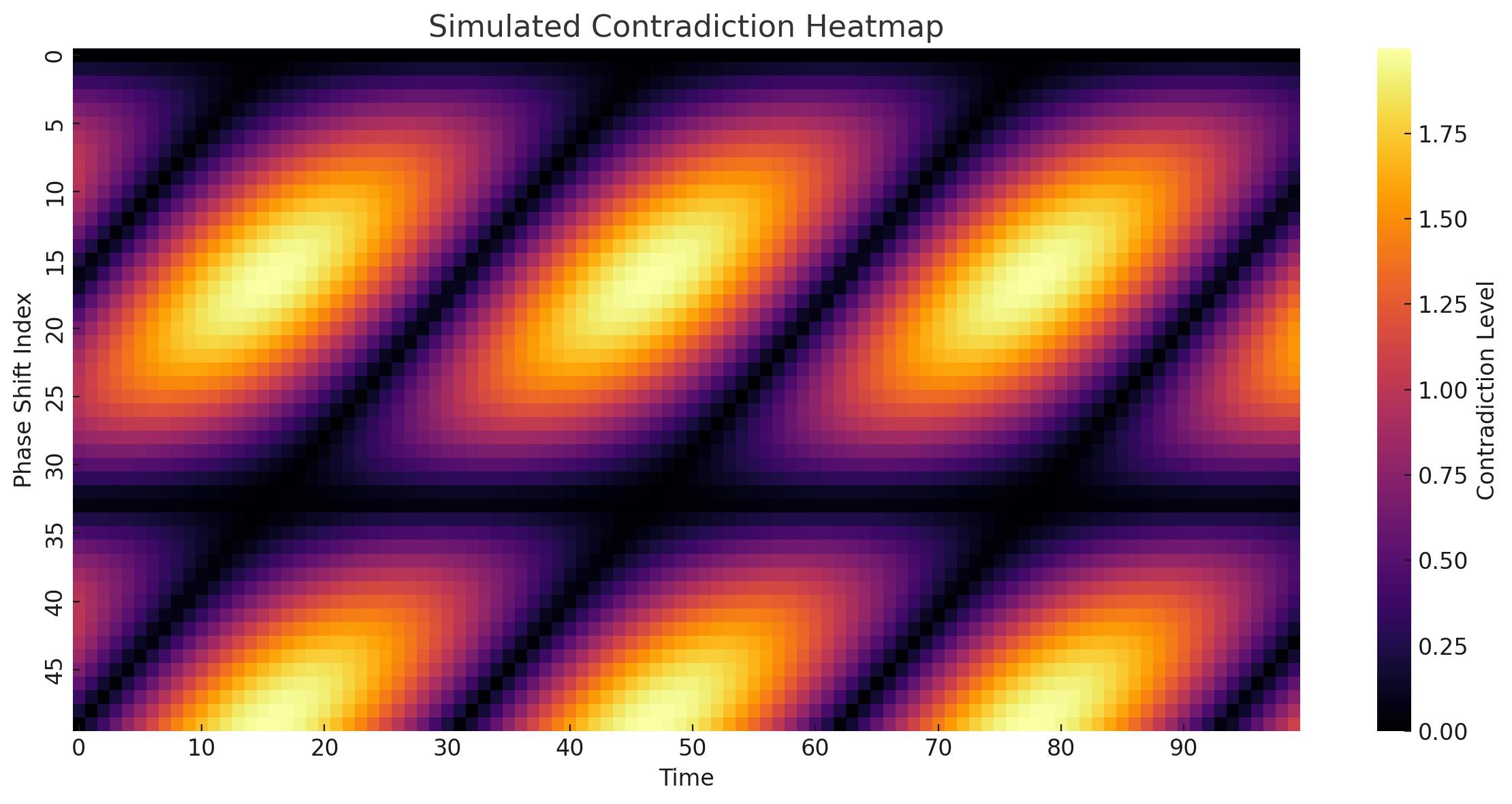
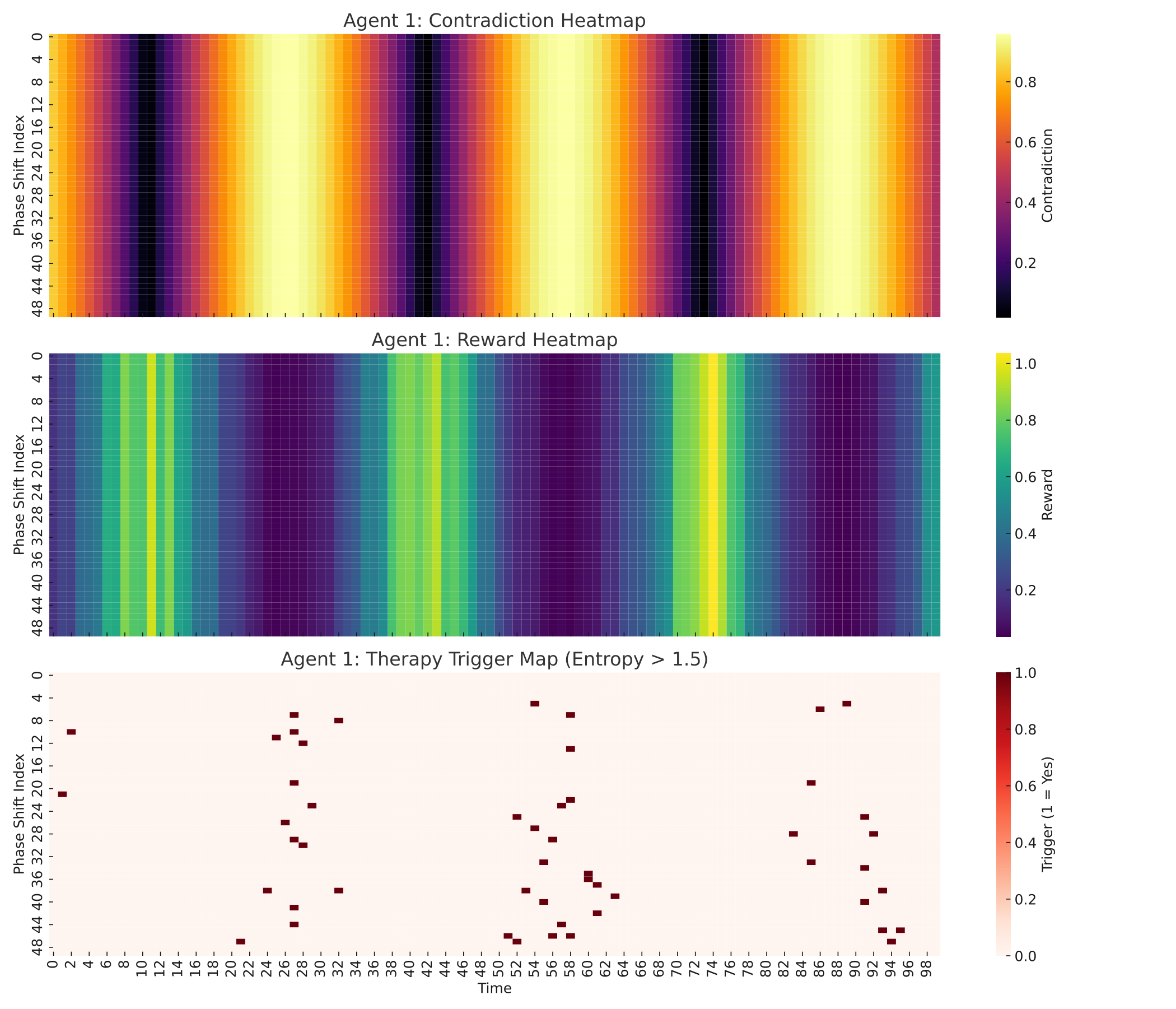
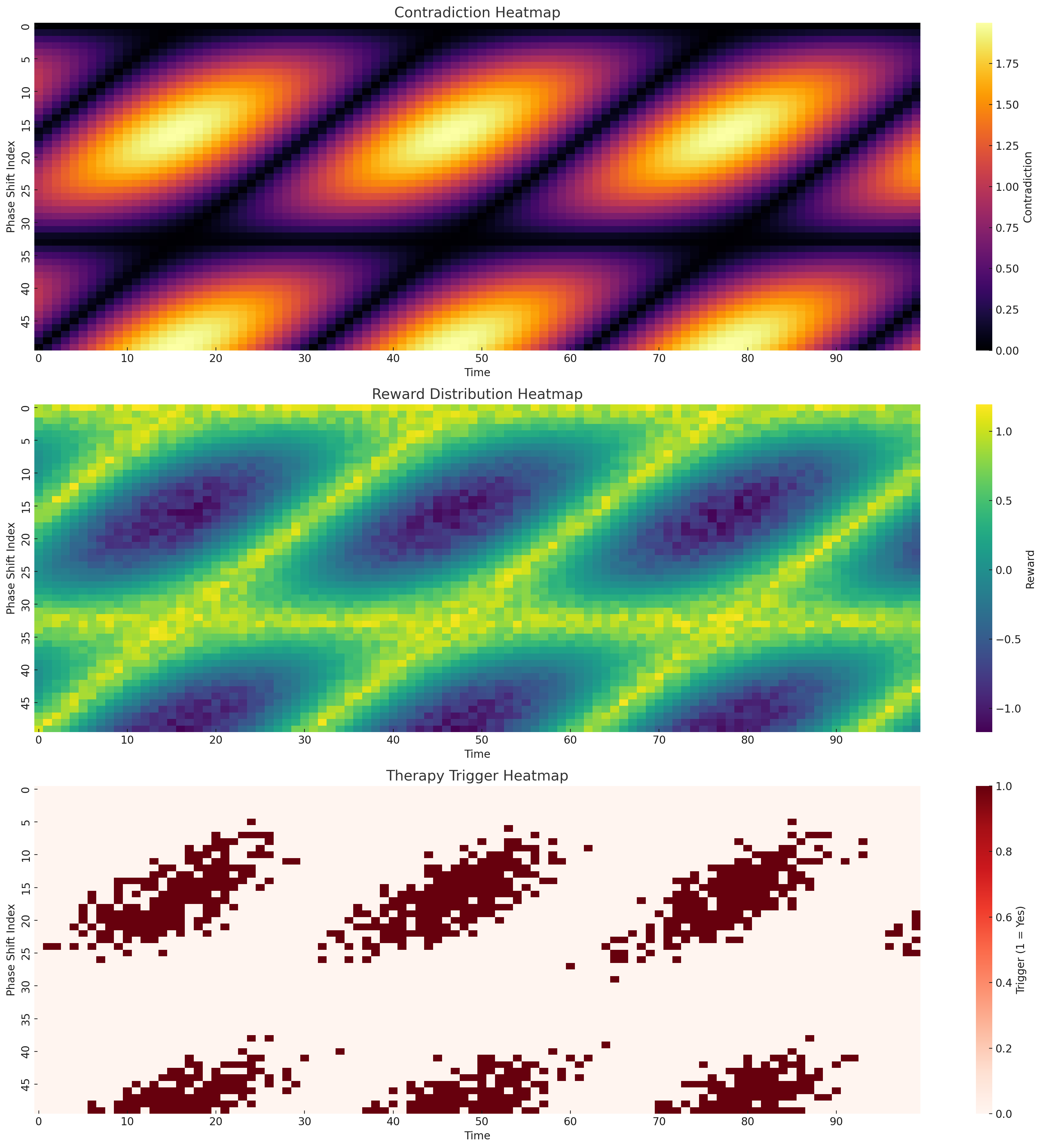
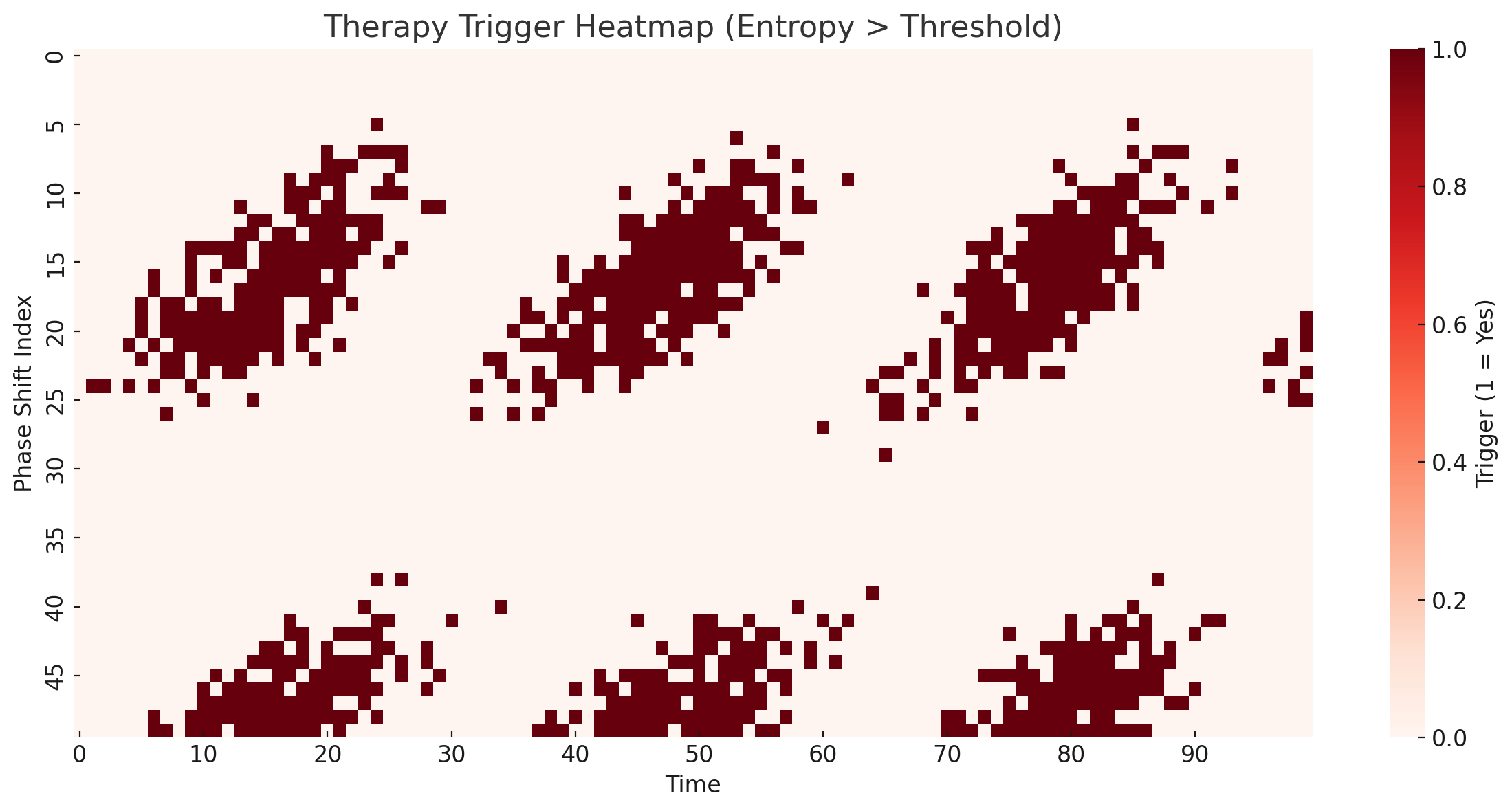
📄

**NUMINA\_AI\_Paper\_v2\_With\_Heatmaps**



**Title:**

NUMINA: A Structure-First AI Model for Symbolic Alignment via Contradiction Minimization and Entropic Feedback

**Abstract**

Modern AI systems often suffer from semantic overfitting and symbolic fixation, producing convincing but incoherent outputs when exposed to contradictory input. We introduce NUMINA, a structure-first symbolic AI framework that minimizes internal contradiction using reinforcement learning and entropy-based feedback. Through a moral-phase simulation, we model contradiction as oscillatory misalignment between an agent’s internal and external symbolic systems. We show how reinforcement learning discovers optimal phase shifts to reduce contradiction. Finally, we demonstrate how entropy signals provide adaptive feedback for symbolic stability.

**1. Introduction**

Language-first AI models, though fluent, are often epistemically brittle. By training on tokens before understanding underlying structure, these models hallucinate, overfit, and collapse under symbolic drift.

NUMINA inverts this paradigm. Inspired by Piaget, Vygotsky, and entropy theory, it learns structural regularities first—then overlays symbolic naming and negotiation. We propose that contradiction itself is a signal—not a failure—and show how resolving contradiction increases symbolic coherence.

**2. Methods**

**2.1 Moral Oscillation Model**

We simulate internal and external morality as oscillating signals:

* Individual: \sin(t + \phi)
* Societal: \cos(t + \pi/4)

Contradiction is defined as:

C(t) = | \sin(t + \phi) - \cos(t + \pi/4) |

**2.2 Entropy as Disorder Proxy**

A moving standard deviation is used to approximate symbolic noise:

H(t) = \text{std}(x\_{t-w:t})

**2.3 Q-Learning Optimization**

A 15-state Q-learning agent explores phase shifts to minimize contradiction and entropy.

Reward function:

R = - \mathbb{E}[C(t)] - \lambda \mathbb{E}[H(t)]

**3. Results**

**Figure 1. Moral Misalignment (Baseline)**

Initial contradiction amplitude is high due to symbolic phase misalignment.

**Figure 2. Learned Alignment**

Reinforcement learning reduces contradiction by shifting individual morality ≈2.3 radians.

**Figure 3. Agent Reward Curve**

Reward signal converges as symbolic contradiction is resolved.

**Figure 4. Entropy Fluctuations**

Post-alignment, the system’s entropy decreases, reflecting symbolic coherence.

**4. Advanced Simulation Results (Agent 1 Dashboard)**

**Figure 5. Contradiction Heatmap**

Heatmap showing internal symbolic contradiction across time and phase shift indices. Darker zones = higher

**Figure 6. Reward Heatmap**

Bright zones correspond to aligned symbolic states with higher cumulative reinforcement reward.

**Figure 7. Therapy Trigger Map**

Binary map highlighting zones where entropy exceeds Z > 1.5, triggering intervention.

**5. Discussion**

The extended simulation shows NUMINA’s symbolic behavior becomes quantifiably more stable as contradiction is reduced. Notably, contradiction heatmaps directly align with areas of low reward and high entropy.

The therapy trigger map confirms that high-entropy regions often coincide with moral dissonance—validating contradiction as a reliable, observable training signal.

These findings support the hypothesis that symbolic coherence can be reinforced by modeling contradiction as an internal alignment challenge—not an error state. This mirrors human development and neural learning, suggesting a scalable framework for symbolic cognition and real-time AI adjustment via entropy-aware interventions.

**6. Conclusion**

NUMINA flips the symbolic script: structure before label, pattern before token. This yields AI agents that adapt, align, and stabilize—even when symbolic drift or contradiction arises. Coupling symbolic phase modeling with entropy-triggered feedback shows promise for building cognitively robust systems.

**7. References**

(stubbed for now)

* Piaget (1954). The Origins of Intelligence in Children
* Shannon (1948). A Mathematical Theory of Communication
* Friston (2006). The Free Energy Principle
* Vygotsky (1962). Thought and Language