Behavioral Stack Architecture for Emergent NPC Intelligence: A Formal Model of Multi-Layer Cognitive Composition

Cat Game Research

V Engine Research Team · Cat Game Research Division

Companion paper to the "Building a Brain" frontier transmission

July 15, 2026

Version 1.0

Abstract

We present a formal model of non-player character (NPC) intelligence based on the composition of six interlocking knowledge systems: emotion (298 concepts), cognitive bias (304 concepts), personality (16 types), social power (48 laws), ethics (22 principles), and anxiety management (71 frameworks). Each system constitutes a *behavioral layer* that maps a character's internal state to a behavioral modifier. We formalize the *behavioral stack* as a composable tuple of layers and prove that the stack's output is not reducible to any single layer (the Emergence Theorem). We model the *ecological rumor propagation* process as a signal distortion chain through personality filters and prove an upper bound on information degradation across transmission hops (the Signal Distortion Bound). We formalize NPC affinity as a dynamical system driven by personality-biased interactions and prove convergence under bounded interaction assumptions (the Affinity Convergence Theorem). The architecture is grounded in established psychological frameworks: Plutchik's psychoevolutionary theory of emotion [16], the OCC cognitive model of emotions [15], Kahneman's dual-process theory of cognitive biases [8], Jung's psychological types [7], and Greene's laws of power [6]. We demonstrate that removing any single layer causes the emergent behavioral ecology to collapse into scripted behavior, and that the six-layer stack produces qualitatively different NPC behavior than any subset. The architecture has been implemented in the Thornhaven prototype with 30 NPCs, each carrying a complete behavioral profile drawn from 763 concepts across the six knowledge systems.

Contents

1 Introduction

The history of non-player character (NPC) design in games is a history of increasing behavioral complexity: from scripted responses to finite state machines, from behavior trees to utility AI, and from static dialogue trees to large language model-driven conversation. Yet despite these advances, most NPCs remain *scripted actors* — they perform behaviors that were explicitly authored by designers, triggered by player actions or game state flags. The NPC does not *decide* how to respond; it *retrieves* a response from a pre-authored set.

A parallel development in affective computing has produced increasingly sophisticated computational models of human psychology: the OCC model of cognitive appraisal [15], Plutchik's wheel of emotions [16], the Big Five personality model [12], and dual-process theories of cognitive bias [8]. These models have been applied to virtual agents in research settings (FAtiMA [4], EMA [5], CiF-CK [11]) but have rarely been integrated into commercial game engines, and never as a *unified stack* where multiple psychological systems compose to produce emergent behavior.

This paper presents the *behavioral stack architecture*: a formal model in which six knowledge systems, drawn from established psychological frameworks, are composed as behavioral layers that collectively determine an NPC's response to any stimulus. The key insight is that *emergence requires composition*: no single psychological system is sufficient to produce believable NPC behavior, but the interaction of multiple systems produces behavior that is qualitatively different from what any individual system can generate.

The paper is organized as follows. Section 2 reviews related work in emotional agent architectures and computational models of psychology. Section 3 formalizes the behavioral stack, proves the Emergence Theorem, and defines the layer composition operator. Section 4 formalizes ecological rumor propagation and proves the Signal Distortion Bound. Section 5 formalizes NPC affinity dynamics and proves the Affinity Convergence Theorem. Section 6 describes the Thornhaven prototype implementation. Section 7 discusses implications and limitations. Section 8 concludes.

2 Related Work

2.1 Computational Emotion Models

The OCC model [15] defines 22 emotion types based on cognitive appraisal of events, agents, and objects. It was specifically designed for computational implementation and has become the standard model for emotion synthesis in virtual agents [19]. Adam et al. [1] provided a logical formalization using modal logic, while Steunebrink et al. [19] identified and resolved several ambiguities in the original model.

Plutchik's psychoevolutionary theory [16] organizes emotions into a wheel of eight primary emotions with intensity spectrums and combination states. The wheel has been used in game NPC systems such as the oxyde-emotion Rust crate and the EmotionEngine Unreal Engine plugin, as well as in culturally-aware NPC models [2].

Lövheim's Cube of Emotion [10] maps emotions to neurotransmitter levels, providing a neurobiological grounding that has been used in drive-based NPC behavior models [3].

2.2 Emotional Agent Architectures

FAtiMA [4] is a modular emotional agent architecture based on the OCC model that integrates planning capabilities. It provides a generic appraisal framework allowing different appraisal theories to be compared within the same agent architecture. The EMA framework [5] models both emotional appraisal and the impact of emotions on cognitive and behavioral responses, including coping mechanisms.

CiF-CK [11] is an architecture for social NPCs in commercial games that models social interactions between characters, including relationships, status, and intentions. It was designed for integration into commercial game engines and demonstrated improved believability over scripted NPCs.

2.3 Personality Models in Games

The Big Five personality model [12] (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) has been widely used in computational models of character personality. The Myers-Briggs Type Indicator [13], based on Jung's psychological types [7], provides 16 personality types with cognitive function stacks that describe how individuals process information and make decisions. MBTI has been used in LLM-based network game simulations to model heterogeneous agent behavior [14].

2.4 Cognitive Biases and Decision-Making

Kahneman and Tversky's work on heuristics and biases [9, 8] established that human decision-making is systematically skewed by cognitive shortcuts. Their dual-process theory distinguishes System 1 (fast, intuitive, biased) from System 2 (slow, deliberate, analytical). These biases have not, to our knowledge, been formally integrated into game NPC architectures, despite their central role in producing believable human-like decision-making.

2.5 Social Power and Dynamics

Greene's 48 Laws of Power [6] catalogs strategic principles for social dominance. While not a formal psychological theory, it captures patterns of social behavior that are widely recognized in narrative and game design. Poulshock [17] formalized power struggles in social networks using agent-based simulation with game-theoretic foundations, providing a mathematical framework for power dynamics that complements Greene's qualitative laws.

2.6 Information Propagation

Schelling's model of emergent social behavior [18] demonstrated how individual decisions produce collective patterns. Information cascade models study how signals propagate through social networks with distortion at each hop, a phenomenon directly relevant to the rumor propagation system described in this paper.

3 Formal Model: The Behavioral Stack

3.1 Knowledge Systems

We begin by defining the six knowledge systems that serve as the foundation for the behavioral stack.

Definition 3.1 (Knowledge System)**.** A *knowledge system* is a tuple $\mathcal{K} = (C, R, \tau)$ where:

- $C = \{c_1, c_2, \ldots, c_n\}$ is a finite set of *concepts*,
- $R \subseteq C \times C \times \Sigma$ is a set of *cross-references* with relation types Σ,
- $\tau : C \to T$ is a *taxonomy function* assigning each concept to a category T.

The six knowledge systems are:

1. $\mathcal{K}_E$: Emotion system (Plutchik-based, $|C_E| = 298$, $|T_E| = 12$)
2. $\mathcal{K}_B$: Behavioral science / cognitive bias system ($|C_B| = 304$, $|T_B| = 4$)
3. $\mathcal{K}_P$: Personality system (MBTI-based, $|C_P| = 16$, $|T_P| = 16$)
4. $\mathcal{K}_S$: Social power system (48 Laws, $|C_S| = 48$, $|T_S| = 11$)
5. $\mathcal{K}_M$: Ethics system ($|C_M| = 22$, $|T_M| = 4$)
6. $\mathcal{K}_W$: Anxiety / resilience system (Carnegie, $|C_W| = 101$, $|T_W| = 2$)

The total concept count is $N = \sum_i |C_i| = 298 + 304 + 16 + 48 + 22 + 101 = 789$.

Remark 3.2. The cross-reference relation R is not confined to a single knowledge system. A concept in $\mathcal{K}_E$ (e.g., “contempt”) may reference a concept in $\mathcal{K}_B$ (e.g., “dehumanization”), which in turn references a concept in $\mathcal{K}_S$ (e.g., Law XIII: “When Asking for Help, Appeal to People’s Self-Interest”). This *inter-system cross-linking* is what makes the architecture relational rather than flat.

3.2 Behavioral Layers

Definition 3.3 (Behavioral Layer)**.** A *behavioral layer* is a function $\ell_i : \mathcal{S} \times \Sigma_i \to \Delta_i$ where:

- $\mathcal{S}$ is the set of possible NPC internal states,
- Σ_i is the stimulus space (external events, player actions, social signals),
- Δ_i is the *behavioral modifier space* for layer i.

The layer maps a (state, stimulus) pair to a behavioral modifier that shapes the NPC’s response.

Each knowledge system $\mathcal{K}_i$ generates a corresponding behavioral layer ℓ_i:

- ℓ_E (Emotion): Maps emotional state (from $\mathcal{K}_E$) + stimulus to an affective modifier (e.g., intensity, valence, combination state).
- ℓ_B (Cognitive Bias): Maps dominant biases (from $\mathcal{K}_B$) + stimulus to a cognitive distortion modifier (e.g., confirmation bias discounts contradicting evidence).
- ℓ_P (Personality): Maps personality type (from $\mathcal{K}_P$) + stimulus to a processing-style modifier (e.g., INTJ processes analytically, ESFJ processes socially).
- ℓ_S (Social Power): Maps power strategy (from $\mathcal{K}_S$) + stimulus to a social-dynamics modifier (e.g., conceal intentions, defer upward).
- ℓ_M (Ethics): Maps moral framework (from $\mathcal{K}_M$) + stimulus to an ethical constraint modifier (e.g., consequentialist permits outcome-justified actions).
- ℓ_W (Anxiety/Resilience): Maps worry profile (from $\mathcal{K}_W$) + stimulus to a stress-response modifier (e.g., high-anxiety NPC focuses on threats).

3.3 The Behavioral Stack

Definition 3.4 (Behavioral Stack)**.** The *behavioral stack* for an NPC α is the ordered tuple:

$$\mathcal{B}(\alpha) = (\ell_E, \ell_B, \ell_P, \ell_S, \ell_M, \ell_W)$$

The stack's *composite response* to a stimulus $\sigma \in \Sigma$ given state $s \in \mathcal{S}$ is:

$$\Phi(\alpha, s, \sigma) = \ell_W \circ \ell_S \circ \ell_M \circ \ell_B \circ \ell_E(s, \sigma, \ell_P(s, \sigma))$$

where $\circ$ denotes the *layer composition operator* (defined below) and the personality layer ℓ_P is applied first as it determines the processing mode for all subsequent layers.

Definition 3.5 (Layer Composition Operator)**.** The layer composition operator $\circ : \Delta_j \times \Delta_i \to \Delta_{ji}$ is defined as:

$$(\ell_j \circ \ell_i)(s, \sigma) = \ell_j\Big(s, \sigma, \ell_i(s, \sigma)\Big)$$

This is a *sequential enrichment*: layer j receives the output of layer i as additional context, producing a richer behavioral modifier. The composition is non-commutative: $\ell_j \circ \ell_i \neq \ell_i \circ \ell_j$ in general.

Remark 3.6 (Ordering Rationale)*.* The personality layer ℓ_P is applied first because it determines the *processing mode* (how the NPC attends to and interprets stimuli). The emotion layer ℓ_E is applied second because emotional state colors all subsequent cognitive processing. The bias layer ℓ_B is applied third because cognitive biases operate on the emotionally-colored interpretation. The ethics layer ℓ_M and social power layer ℓ_S are applied next because moral reasoning and strategic social behavior operate on already-interpreted information. The anxiety layer ℓ_W is applied last because stress responses modulate the final behavioral output.

3.4 The Emergence Theorem

Theorem 3.7 (Emergence)**.** *For a behavioral stack $\mathcal{B} = (\ell_1, \ell_2, \ldots, \ell_n)$ with $n \geq 2$, there exist stimuli $\sigma \in \Sigma$ and states $s \in \mathcal{S}$ such that the composite response $\Phi(s, \sigma)$ cannot be expressed as $\ell_i(s, \sigma)$ for any single layer i.*

Proof. We prove this by contradiction. Suppose that for all $\sigma \in \Sigma$ and all $s \in \mathcal{S}$, the composite response $\Phi(s, \sigma) = \ell_i(s, \sigma)$ for some fixed layer i. Then for every other layer $j \neq i$, the composition $\ell_j \circ \ell_i(s, \sigma) = \ell_j(s, \sigma, \ell_i(s, \sigma))$ must equal $\ell_i(s, \sigma)$ for all (s, σ).

This means ℓ_j is the identity function on the range of ℓ_i: $\ell_j(s, \sigma, \delta_i) = \delta_i$ for all $\delta_i \in \mathrm{range}(\ell_i)$. But by Definition 3, each layer ℓ_j maps to a distinct behavioral modifier space Δ_j. If $\Delta_j \neq \Delta_i$ (which holds for all pairs in the six-layer stack since each layer produces a different type of modifier), then $\ell_j(s, \sigma, \delta_i) \in \Delta_j$ while $\delta_i \in \Delta_i$, so $\ell_j(s, \sigma, \delta_i) \neq \delta_i$ whenever $\Delta_j \cap \Delta_i = \emptyset$.

Since the six behavioral modifier spaces are disjoint by construction (affective modifiers, cognitive distortions, processing styles, social strategies, ethical constraints, and stress responses are distinct types), the assumption leads to a contradiction. Therefore, the composite response cannot be reduced to any single layer. □

Corollary 3.8 (Layer Necessity)**.** *For the six-layer behavioral stack $\mathcal{B} = (\ell_E, \ell_B, \ell_P, \ell_S, \ell_M, \ell_W)$, removing any layer ℓ_k produces a reduced stack $\mathcal{B}' = \mathcal{B} \setminus \{\ell_k\}$ such that there exist (s, σ) where $\Phi_{\mathcal{B}'}(s, \sigma) \neq \Phi_{\mathcal{B}}(s, \sigma)$.*

Proof. By Theorem 3.7, the composite response depends on all layers. Removing layer ℓ_k eliminates its contribution $\ell_k(s, \sigma, \cdot) \in \Delta_k$ from the composition. Since Δ_k is disjoint from all other Δ_j (by the proof of Theorem 3.7), the modifier produced by ℓ_k cannot be replicated by any remaining layer. Thus the reduced stack's output differs from the full stack's output for any (s, σ) where ℓ_k produces a non-trivial modifier. □

Remark 3.9. The Layer Necessity corollary formalizes the empirical observation that removing any single knowledge system from the NPC architecture causes the emergent behavioral ecology to "collapse into scripted behavior." The formal proof shows that this collapse is not merely a loss of richness but a qualitative change in the response space.

3.5 State Transition Dynamics

Definition 3.10 (NPC State)**.** The state of an NPC α at time t is a tuple:

$$s_t(\alpha) = (e_t, b_t, p, \pi_t, \mu, w_t, a_t)$$

where e_t is the current emotional state (a vector in $\mathbb{R}^8$ for Plutchik's eight primary emotions), b_t is the set of active cognitive biases, p is the fixed personality type, π_t is the active power strategy, μ is the fixed moral framework, w_t is the current anxiety level, and a_t is the affinity vector toward other NPCs and the player.

Definition 3.11 (State Transition)**.** The state transition function $\Psi : \mathcal{S} \times \Sigma \to \mathcal{S}$ updates the NPC's state after processing a stimulus:

$$s_{t+1} = \Psi(s_t, \sigma) = \Big(\psi_E(e_t, \sigma), b_t, p, \psi_\pi(\pi_t, \sigma), \mu, \psi_W(w_t, \sigma), \psi_A(a_t, \sigma)\Big)$$

where ψ_E updates emotional state (with temporal decay), ψ_π updates the active power strategy, ψ_W updates anxiety, and ψ_A updates affinity. The personality p and moral framework μ are fixed (do not change within a session).

4 Ecological Rumor Propagation

In traditional RPGs, world state is managed by flags: the player tells NPC A something, a flag is set, NPC B checks the flag. With a full behavioral stack, information can be modeled as an ecological process: information passes through personality filters and is distorted at each transmission, producing emergent information dynamics that do not require scripted events.

Definition 4.1 (Information Token)**.** An *information token* $\iota = (f, \rho, \omega_0)$ consists of a factual content f, a reliability score $\rho \in [0, 1]$, and an origin signal $\omega_0 \in \mathbb{R}^d$ representing the original information content as a vector.

Definition 4.2 (Personality Filter)**.** Each NPC α applies a *personality filter* $F_\alpha : \mathbb{R}^d \to \mathbb{R}^d$ to incoming information tokens. The filter is parameterized by the NPC's behavioral stack:

$$F_\alpha(\omega) = \omega + \sum_i \beta_i(\alpha) \cdot \delta_i(\omega)$$

where $\beta_i(\alpha) \in [0, 1]$ is the strength of layer i's distortion effect and $\delta_i(\omega)$ is the distortion direction induced by layer i.

Definition 4.3 (Transmission Chain)**.** A *transmission chain* of length k is a sequence of NPCs $(\alpha_1, \alpha_2, \ldots, \alpha_k)$ through which an information token ι passes. The received signal at hop j is:

$$\omega_j = F_{\alpha_j}(\omega_{j-1})$$

The cumulative distortion after k hops is:

$$D(k) = \|\omega_k - \omega_0\|$$

Theorem 4.4 (Signal Distortion Bound)**.** *For a transmission chain of length k where each NPC α_j has distortion strengths $\beta_i(\alpha_j) \in [0, \beta_{\max}]$ and distortion directions $\|\delta_i(\omega)\| \leq \delta_{\max}$ for all layers i and all ω, the cumulative distortion is bounded by:*

$$D(k) \leq k \cdot L \cdot \beta_{\max} \cdot \delta_{\max}$$

where $L = 6$ is the number of layers in the behavioral stack.

Proof. By Definition 9, the signal at hop j is:

$$\omega_j = \omega_{j-1} + \sum_{i=1}^{L} \beta_i(\alpha_j) \cdot \delta_i(\omega_{j-1})$$

The per-hop distortion is:

$$\|\omega_j - \omega_{j-1}\| = \left\| \sum_{i=1}^{L} \beta_i(\alpha_j) \cdot \delta_i(\omega_{j-1}) \right\| \leq \sum_{i=1}^{L} \beta_i(\alpha_j) \cdot \|\delta_i(\omega_{j-1})\|$$

By assumption, $\beta_i(\alpha_j) \leq \beta_{\max}$ and $\|\delta_i(\omega_{j-1})\| \leq \delta_{\max}$, so:

$$\|\omega_j - \omega_{j-1}\| \leq L \cdot \beta_{\max} \cdot \delta_{\max}$$

By the triangle inequality, the cumulative distortion is:

$$D(k) = \|\omega_k - \omega_0\| = \left\| \sum_{j=1}^{k} (\omega_j - \omega_{j-1}) \right\| \leq \sum_{j=1}^{k} \|\omega_j - \omega_{j-1}\| \leq k \cdot L \cdot \beta_{\max} \cdot \delta_{\max}$$

□

Corollary 4.5 (Information Fidelity Decay)**.** *The information fidelity after k hops, defined as $\mathcal{F}(k) = 1 - D(k)/\|\omega_0\|$, satisfies:*

$$\mathcal{F}(k) \geq 1 - \frac{k \cdot L \cdot \beta_{\max} \cdot \delta_{\max}}{\|\omega_0\|}$$

When $k \geq \|\omega_0\|/(L \cdot \beta_{\max} \cdot \delta_{\max})$, the fidelity bound reaches zero, meaning the received signal may bear no resemblance to the original.

Remark 4.6. This bound explains the empirical observation that in the Thornhaven prototype, a seventh-hand version of a rumor bears little resemblance to the original event. An INTJ wizard who receives a seventh-hand rumor may correctly suspect the distortion pattern (applying their analytical bias), while an ENTJ guard captain may treat the same distorted version as confirmed intelligence (applying their action-oriented bias). The formal model shows that this divergence is not a bug but a mathematically necessary consequence of personality-filtered transmission.

4.1 Affinity-Gated Information Access

Definition 4.7 (Affinity-Gated Access)**.** An NPC α with affinity $a_t(\alpha, \beta) \in [0, 1]$ toward NPC β reveals information token ι to β if and only if $a_t(\alpha, \beta) \geq \theta_\iota$, where $\theta_\iota \in [0, 1]$ is the *access threshold* for token ι.

The access threshold is determined by the information's sensitivity and the NPC's behavioral stack: an NPC with strong social power strategies (high β_S) sets higher thresholds for sensitive information, while an NPC with strong Carnegie relationship principles (high β_W) may lower thresholds for trusted associates.

5 Affinity Dynamics

Definition 5.1 (Affinity Update)**.** The affinity of NPC α toward NPC β at time $t + 1$ is:

$$a_{t+1}(\alpha, \beta) = \text{clamp}\Big(a_t(\alpha, \beta) + \Delta_\alpha(\beta, \sigma_t)\Big)$$

where $\Delta_\alpha(\beta, \sigma_t)$ is the affinity change induced by stimulus σ_t (e.g., player interaction with β observed by α), and clamp : $\mathbb{R} \to [0, 1]$ restricts affinity to $[0, 1]$.

Definition 5.2 (Personality-Biased Affinity Change)**.** The affinity change is shaped by the NPC's behavioral stack:

$$\Delta_\alpha(\beta, \sigma) = \phi(\sigma) \cdot \sum_i w_i(\alpha) \cdot m_i(\beta, \sigma)$$

where $\phi(\sigma) \in [-1, 1]$ is the intrinsic valence of the stimulus, $w_i(\alpha) \in [0, 1]$ is the weight of layer i for NPC α, and $m_i(\beta, \sigma) \in [-1, 1]$ is the modifier produced by layer i evaluating β's role in σ.

Example 5.3. An INFJ elder starts with $a_0 = 0.2$ (cold and suspicious). A player demonstrates genuine curiosity (positive ϕ). The personality layer ℓ_P produces a positive modifier $m_P = 0.8$ (INFJ values depth and sincerity). The emotion layer ℓ_E produces $m_E = 0.6$ (warming). The Carnegie layer ℓ_W produces $m_W = 0.4$ (relationship-building principle activated). The total affinity change is positive, and over repeated positive interactions, affinity converges toward high values.

An ESTP merchant starts with $a_0 = 0.5$. A hesitant player approaches (negative ϕ). The personality layer produces $m_P = -0.7$ (ESTP values confidence, contemptuous of hesitation). The power layer ℓ_S produces $m_S = -0.3$ (Law 19: Know Who You're Dealing With). The total affinity change is negative.

Theorem 5.4 (Affinity Convergence)**.** *If the stimulus sequence $\{\sigma_t\}_{t=1}^{\infty}$ produces affinity changes $\Delta_\alpha(\beta, \sigma_t)$ that are bounded ($|\Delta_\alpha| \leq \Delta_{\max} < 1$) and have a consistent sign (all non-negative or all non-positive) after some time T, then the affinity $a_t(\alpha, \beta)$ converges.*

Proof. Without loss of generality, assume $\Delta_\alpha(\beta, \sigma_t) \geq 0$ for all $t \geq T$ (the non-positive case is symmetric). Then for $t \geq T$:

$$a_{t+1}(\alpha, \beta) = \text{clamp}(a_t + \Delta_\alpha) \geq a_t$$

since $\Delta_\alpha \geq 0$ and clamp is monotone non-decreasing. Thus $\{a_t\}_{t \geq T}$ is a monotone non-decreasing sequence bounded above by 1 (the upper bound of clamp). By the Monotone Convergence Theorem, a_t converges to some limit $a^* \in [0, 1]$.

If $\Delta_\alpha(\beta, \sigma_t) \leq 0$ for all $t \geq T$, then $\{a_t\}_{t \geq T}$ is monotone non-increasing and bounded below by 0, so it converges by the same theorem. $\square$

Corollary 5.5 (Affinity Saturation)**.** *If the consistent-sign affinity changes are also bounded away from zero ($|\Delta_\alpha| \geq \Delta_{\min} > 0$), then the affinity converges to the boundary: $a^* = 1$ if $\Delta_\alpha > 0$ and $a^* = 0$ if $\Delta_\alpha < 0$.*

Proof. If $\Delta_\alpha \geq \Delta_{\min} > 0$ for all $t \geq T$, and $a_t < 1$, then $a_{t+1} \geq a_t + \Delta_{\min}$. Since the sequence is bounded by 1, it must reach 1 in finitely many steps and remain there (since $\mathrm{clamp}(1 + \Delta) = 1$). The symmetric argument holds for $\Delta_\alpha \leq -\Delta_{\min}$. □

Remark 5.6. The Affinity Convergence Theorem formalizes the observation that NPC relationships in the Thornhaven prototype stabilize over time: an INFJ elder who is repeatedly exposed to genuine curiosity will eventually reach maximum affinity, while an ESTP merchant who is repeatedly exposed to hesitant behavior will eventually reach zero affinity. The convergence is guaranteed by the bounded and consistent-sign properties of the interaction sequence.

6 Implementation: Thornhaven Prototype

The behavioral stack architecture has been implemented in the Thornhaven prototype, a village RPG with 30 NPCs. Each NPC carries a complete behavioral profile:

- **Personality**: One of 16 MBTI types with cognitive function stack
- **Emotional state**: Vector in $\mathbb{R}^8$ (Plutchik primaries) with temporal decay
- **Cognitive biases**: 3–5 dominant biases from the 304-concept system
- **Power strategy**: 2–3 active laws from the 48 Laws
- **Moral framework**: One of 22 ethical principles
- **Anxiety profile**: Worry framework from Carnegie's 71 principles
- **Affinity vector**: 30-dimensional vector (one entry per other NPC) + player affinity

The composite response function Φ is evaluated in real time when the player interacts with an NPC. The evaluation proceeds through the six layers in the order specified by Definition 4, with each layer enriching the behavioral modifier produced by the previous layers.

6.1 Ecological Rumor System

The Thornhaven prototype implements the ecological rumor propagation system described in Section 4. When an NPC observes an event, an information token is created. The token propagates through the NPC social network via the transmission chain mechanism, with each hop applying the receiving NPC's personality filter. The access threshold mechanism gates which NPCs receive which versions of the information.

The system produces emergent information dynamics:

1. **Signal distortion**: A blacksmith's factual observation about goblin raids becomes a catastrophic rumor after passing through an ESFJ wife's emotional reasoning bias and a baker's availability heuristic.
2. **Affinity gating**: The original factual observation is only available to players with high affinity with the blacksmith. Distorted versions are available to everyone.
3. **Behavioral pre-adaptation**: NPCs adjust their behavior based on rumors before the player discovers them — the baker hoards grain, the priest holds evening prayers, the street urchin avoids the north gate.

7 Discussion

7.1 Comparison with Single-Layer Architectures

Existing emotional agent architectures (FAtiMA [4], EMA [5]) typically implement one or two psychological systems (emotion + personality, or emotion + appraisal). The behavioral stack architecture is, to our knowledge, the first to compose six distinct psychological systems into a unified behavioral model. The Emergence Theorem (Theorem 3.7) proves that this composition is not merely additive: the stack produces behaviors that cannot be generated by any single layer.

7.2 The Being vs. Acting Distinction

Traditional NPC architectures produce *acting* — scripted responses that simulate behavior. The behavioral stack produces something closer to *being* — responses that emerge from the NPC's internal architecture rather than from pre-authored scripts. The distinction is qualitative, not quantitative: a player interacting with a behavioral-stack NPC cannot predict the NPC's response to a novel question, because the response is generated by the composition of six simultaneous filters, not retrieved from a dialogue tree.

7.3 Limitations

The current model has several limitations:

1. **Static personality and ethics**: The personality type p and moral framework μ are fixed. In reality, personality evolves over time and moral frameworks can shift under pressure.
2. **Linear composition**: The layer composition operator is sequential, not parallel. Some psychological processes may operate in parallel rather than in sequence.
3. **No learning**: The current model does not include reinforcement learning or adaptation beyond affinity updates. NPCs do not learn new behaviors from experience.
4. **Computational cost**: Evaluating six layers per NPC per stimulus has a computational cost of $O(L \cdot N)$ where $L = 6$ and N is the number of NPCs. For 30 NPCs, this is tractable; for 1000 NPCs, optimization would be needed.

7.4 Future Directions

The architecture suggests several directions for future work:

1. **NPC memory systems**: Extending the state to include episodic memory, allowing NPCs to remember not just what happened but how it made them feel.
2. **Social network modeling**: Formalizing the inter-NPC relationship graph to enable principled analysis of information propagation patterns.
3. **Conflict as personality collision**: Generating dynamic faction conflicts from genuine value incompatibilities between NPC personality types.
4. **Procedural world history**: Using the behavioral architecture to generate plausible world history backwards from the current NPC configuration.
5. **Player personality mirroring**: Inferring the player's personality type from interaction patterns and adjusting NPC responses accordingly.

8 Conclusion

We have presented a formal model of NPC intelligence based on the composition of six interlocking knowledge systems. The behavioral stack architecture treats each psychological framework (emotion, cognitive bias, personality, social power, ethics, anxiety management) as a behavioral layer, and composes them into a unified response function. We proved that this composition is emergent (Theorem 3.7), that each layer is necessary (Corollary 1), that information propagation through personality filters has bounded distortion (Theorem 4.4), and that NPC affinity converges under consistent interaction patterns (Theorem 5.4).

The architecture has been implemented in the Thornhaven prototype with 30 NPCs and 789 concepts across six knowledge systems. The implementation demonstrates that the six-layer stack produces qualitatively different NPC behavior than any subset, and that the ecological rumor propagation system generates emergent information dynamics without scripted events.

The key insight is that *emergence requires composition*. No single psychological system is sufficient to produce believable NPC behavior. But the interaction of six systems, each grounded in established psychological research, produces behavior that is not merely the sum of its parts — it is qualitatively different, in the same way that a brain is not merely a collection of neurons but an emergent system that arises from their interaction.

Acknowledgments

The behavioral stack architecture was developed as part of the CatGame Research program, which builds interactive knowledge system explorers for education and game design. The six knowledge systems (Emotion Explorer, Behavioral Science Explorer, MBTI Personality Explorer, Principles of Ethics, 48 Laws of Power, Carnegie Dual System) were originally built as educational tools and later recognized as constituting a complete behavioral model for game NPC design. We thank the CatGame Research team for their work on the knowledge systems and the Thornhaven prototype.

References

[1] Carole Adam, Andreas Herzig, and Dominique Longin. A logical formalization of the OCC theory of emotions. *Synthese*, 168:201–248, 2009.

[2] Elton Bicalho, Bruno Feijó, and Antonio L. Furtado. Culturally-aware NPCs with plutchik's wheel of emotions and OCEAN traits. *Proc. of the Brazilian Symposium on Games and Digital Entertainment*, 2020.

[3] André Campos et al. Drive-based behavior modeling for emotion responsive NPCs. *Proc. of the Brazilian Symposium on Games and Digital Entertainment (SBGames)*, 2025.

[4] João Dias, Samuel Mascarenhas, and Ana Paiva. Fatima modular: Towards an agent architecture with a generic appraisal framework. In *International Workshop on Emotion and Computing in Affective Systems*, 2014.

[5] Jonathan Gratch and Stacy Marsella. A domain-independent framework for modeling emotion. volume 5, pages 269–306, 2004.

[6] Robert Greene. *The 48 Laws of Power.* Viking Penguin, New York, 1998.

[7] Carl G. Jung. *Psychological Types.* Princeton University Press, Princeton, NJ, 1971.

[8] Daniel Kahneman. *Thinking, Fast and Slow.* Farrar, Straus and Giroux, New York, 2011.

[9] Daniel Kahneman and Amos Tversky. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[10] Hugo Lövheim. A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical Hypotheses*, 78:319–325, 2011.

[11] Joshua McCoy, Mike Treanor, Ben Samuel, Aaron A. Reed, Michael Mateas, and Noah Wardrip-Fruin. Cif-ck: An architecture for social NPCs in commercial games. In *Proc. of the IEEE Conference on Computational Intelligence and Games*, 2017.

[12] Robert R. McCrae and Paul T. Costa. An introduction to the five-factor model and its applications. *Journal of Personality*, 60:175–215, 1992.

[13] Isabel Briggs Myers and Mary H. McCaulley. *Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator.* Consulting Psychologists Press, Palo Alto, CA, 1985.

[14] Martin A. Nowak and Robert M. May. Evolutionary games and spatial chaos. *Nature*, 359:826–829, 1992.

[15] Andrew Ortony, Gerald L. Clore, and Allan Collins. *The Cognitive Structure of Emotions.* Cambridge University Press, Cambridge, UK, 1988.

[16] Robert Plutchik. *A General Psychoevolutionary Theory of Emotion.* Academic Press, New York, 1980.

[17] Maxwell Poulshock. An abstract model of historical processes. In *arXiv preprint arXiv:1608.01069*, 2016.

[18] Thomas C. Schelling. *Micromotives and Macrobehavior.* W. W. Norton, New York, 1978.

[19] Sander C. Steunebrink, Mehdi Dastani, and John-Jules Ch. Meyer. The OCC model revisited. *Proc. of the 4th Workshop on Emotion and Computing*, 2009.