Emergent Emotional Appraisal in Generative Agents via Predictive World Modeling

Prem Babu Kanaparthi
Rochester Institute of Technology
Rochester, New York, USA
prem.b.kanaparthi@gmail.com

Figure 1: Ghost Town simulation environment with emergent emotion engine. The simulation tests whether emotional appraisal patterns can emerge from predictive world modeling without explicit emotion rules or supervision.

Abstract

We investigate whether behavioral analogs to canonical emotional appraisal patterns can arise as a consequence of predictive world modeling in generative agent systems, without explicit emotion rules, reward shaping, or labeled affect. We introduce Ghost Town, a 12-agent survival simulation, and conduct a five-condition controlled comparison spanning 8 random seeds, 6 scenario variants, and $N = 205{,}940$ agent-step records. The proposed method, Condition D (predictive world modeling), trains a dual-head MLP to jointly predict agent actions and 12 binary future world events; emotion quantities are derived entirely post hoc from prediction confidence and are never optimized during training. Three behavioral signatures corresponding to OCC cognitive appraisal constructs [5] emerge in Condition D without supervision: fear-driven shelter-seeking (99.0% vs. 52.7% under baseline, $p = 3.3 \times 10^{-113}$); suspicion decay as a function of time since betrayal (64.6% $\rightarrow$ 35.2%, $p = 8.7 \times 10^{-142}$); and structured latent dimensions aligning with appraisal axes ($\chi^2 \geq 52.6$, $p < 0.01$). Condition D outperforms all four baselines on survival, cooperative rescue, and agent trust. As a preliminary cross-population check, resting grief activation (25.5%) aligns within 0.5 pp of Gallup 2024 global sadness prevalence (26%, $N = 145{,}000$), with no calibration to human data. These results are consistent with the predictive processing account [3]: emotion-like behavioral structure may arise as a compressed consequence of world-event prediction rather than as an explicitly engineered module.

CCS Concepts

• **Computing methodologies** → **Multi-agent systems**; *Reinforcement learning*; *Cognitive science.*

Keywords

generative agents, emotional appraisal, predictive processing, OCC theory, multi-agent simulation, emergent behavior

1 Introduction

Emotions in biological agents are not decorative overlays on cognition. Fear directs attention toward threats before conscious reasoning can act; grief restructures social priorities after loss; and suspicion calibrates trust over repeated interactions. If artificial agents are to produce human-like social dynamics, rather than merely mimicking surface behavior, they require internal states that serve these functional roles. The central question is therefore: *where do such states originate?*

Park et al. [8] introduced Generative Agents, a landmark system in which 25 LLM-driven agents inhabit Smallville, a sandbox world reminiscent of *The Sims*, and exhibit emergent social behaviors such as planning daily routines, gossiping, forming opinions, and organizing events (e.g., a Valentine's Day party emerging from a single seed suggestion). The architecture relies on a *memory stream*, a natural-language database of agent experiences, which is retrieved to drive planning, reflection, and action via GPT-3.5-turbo. Agents report emotional states within this narrative (e.g., "I felt anxious today"), but these are generated as part of text output rather than produced by an explicit computational mechanism. Consequently, the *origin* of emotion as a process that can be designed, analyzed, or ablated remains unaddressed. This paper directly addresses this gap.

Three standard approaches appear in the AI literature:

(1) **Hand-coded emotion rules.** For example, "if ghost visible, fear += 0.3". These perform well in anticipated scenarios but are brittle under novel conditions.
(2) **Imitation of emotional behavior.** Behavioral cloning on trajectories from rule-based agents. While surface behavior is reproduced, internal representations remain shallow.
(3) **Modeling emotion dynamics.** Learning transitions between emotional states over time. This yields richer structure but remains anchored to predefined dimensions.

We propose a fourth approach grounded in predictive processing theory [2, 3]. Under predictive processing, emotions function as compressed internal representations of anticipated consequences not as outputs of a dedicated affective module [1]. If an agent must accurately predict "will a ghost appear in 3 steps?" or "will an ally die in 5 steps?", the latent features most useful for those predictions will organize around danger, mortality, and social threat. These features correspond functionally to OCC appraisal dimensions. Crucially, the model receives no explicit emotion labels; any emotion-like structure that arises is a post hoc *interpretable consequence* of optimizing the predictive objective, not a design target.

We construct a five-condition empirical testbed, the Ghost Town simulation, and provide comprehensive experimental evidence, including statistical analysis over 205,940 agent-step records, agent interaction logs, affect timelines, and training data samples.

Contributions.

- A five-condition controlled comparison of emotion architectures in a survival simulation (12 agents, 72 steps, 6 scenarios, 8 seeds).
- Statistical evidence that Condition D independently recovers three OCC appraisal signatures without emotion labels ($p < 10^{-100}$).
- An interpretability analysis demonstrating that latent dimensions align with OCC appraisal structure.
- A preliminary cross-population calibration against Gallup 2024 ($N = 145,000$) global emotion prevalence data.

2 Background

2.1 OCC Cognitive Appraisal Theory

The dominant computational theory of emotion [5] proposes that emotions arise from cognitive appraisal, i.e., the evaluation of events relative to goals, standards, and attitudes. Fear emerges when a situation is appraised as threatening to goals and beyond coping capacity. Grief arises when a loss is appraised as significant and irreversible. Suspicion arises when another agent's intentions are appraised as hostile; this appraisal weakens as memory of the triggering event decays.

Three behavioral signatures operationalize these appraisals:

- **Fear → Shelter**: agents are more likely to seek safe locations when a threat is visible.
- **Earned Distrust**: refusal rate increases with betrayal count.
- **Suspicion Decay**: forgiveness rate increases as time since betrayal grows.

Our central hypothesis is that a model trained solely on future event prediction can recover all three signatures without explicit supervision.

2.2 Predictive Processing and Constructed Emotion

Predictive processing [2, 3] posits that the brain functions as a prediction system: it continuously models the world, generates predictions, and updates internal representations in response to prediction errors. Within this framework, what are commonly described as "emotions" can be viewed as compressed summaries of anticipated outcomes, rather than outputs of a separate affective module. Barrett [1] extends this perspective by arguing that emotional categories are constructed from interoceptive predictions rather than being innately specified.

Condition D operationalizes this idea computationally. Specifically, the model's "fear" corresponds to its estimate of $P(\texttt{ghost\_nearby\_t3})$, and its "grief" corresponds to a function of $P(\texttt{nearby\_death\_t5})$. Importantly, no explicit emotion labels are provided during training.

3 The Ghost Town Simulation

3.1 Environment Design

Ghost Town is a 12-agent survival simulation designed to provide controlled and repeatable conditions for studying emergent emotional behavior. Unlike the Smallville social sandbox of Park et al. [8], which focuses on everyday social life (e.g., work, gossip, coordination), Ghost Town introduces life-or-death threat, resource scarcity, and social betrayal conditions that strongly engage fear, grief, and suspicion appraisal mechanisms.

Table 1: Agent roster. Traits are drawn uniformly at agent creation; initial ties are set by relationship label (household > 0.5, neighbors ≈ 0.1–0.3).

Name	Role	Courage	Empathy
Alma Ward	Town Steward	0.72	0.88
Dr. Mira Chen	Medic	0.61	0.91
Owen Pike	Farmhand	0.55	0.72
Sheriff Elias Boone	Sheriff	0.82	0.65
Nora Vale	Teacher	0.68	0.84
Rosa Mercer	Innkeeper	0.59	0.77
Caleb Dunn	Line Cook	0.52	0.63
Gideon Marsh	Mechanic	0.74	0.58
Wren Holloway	Apprentice	0.45	0.80
June Carter	Herbalist	0.63	0.86
Silas Reed	Watchman	0.78	0.60
Ivy Hart	Seamstress	0.50	0.82

Agents navigate a shared grid world (40×25 tiles) over 72 discrete steps representing three simulated days. Each step corresponds to approximately 30 minutes of in-world time.

3.2 Agent Roster

Twelve agents populate every run with fixed names, roles, traits, and initial social ties (Table 1).

3.3 Scenarios

Six scenario variants modulate the threat environment across three categories: existential (ghost and death), resource (scarcity), and social (betrayal and crowding). Table 2 summarizes each scenario and the primary OCC signature it is designed to elicit.

All primary statistical results use `standard_night` with 8 seeds (seeds 0–7). Generalization results (Section 6.6) aggregate all six scenarios.

4 Five Emotion-Engine Conditions

Each condition provides agents with a distinct cognitive architecture for computing internal emotional state from observations. All other simulation components, including world mechanics, pathfinding, supply economy, social graph, and dialogue, are identical across conditions (Table 3).

4.1 Baseline (Condition 0): No Emotion

The Baseline condition contains no emotion engine. Agents act using fixed priority heuristics: seek shelter at night, gather supplies during the day, and help allies when health falls below a threshold. No internal state is tracked beyond health, stamina, and inventory. This condition establishes the lower bound for all comparisons. Observed shelter-seeking rates reflect planner behavior rather than emotional modulation.

Table 2: Six scenario variants. "Signature" denotes the primary OCC appraisal dimension activated by each scenario's mechanics.

Scenario	Signature	Description
`standard_night`	Fear	Baseline ghost pressure with 1 ghost per night. Eight primary seeds. Food and shelter are available. Used for all main results.
`high_ghost_pressure`	Fear (extreme)	Three simultaneous ghosts per night phase. Tests fear calibration under saturation.
`storm_scarcity`	Stress + Grief	Rolling storm reduces food supply by 70%. Tests stress and grief under material threat.
`ally_death`	Grief	June Carter dies in a scripted ghost strike at step 1. Tests grief activation and behavioral reorganization after loss.
`betrayal_refusal`	Suspicion	Selected pairs are initialized as rivals (tie -0.68 to -0.95). Tests suspicion accumulation and decay over 72 steps.
`crowded_shelter`	Fear + Suspicion	Safe building capacity is halved (6 slots for 12 agents). Tests cooperative versus competitive shelter allocation.

Table 3: Five emotion-engine conditions. Only the emotion architecture differs. All conditions share the same world, agents, and random seeds.

Condition	Architecture
Baseline (0)	No emotion system; priority heuristics only.
Cond A	Hand-coded OCC appraisal rules.
Cond B	Behavioral cloning on Condition A trajectories.
Cond C	Emotion dynamics model (latent transition).
Cond D	**Predictive world modeling (proposed).**

4.2 Condition A: Hand-Coded OCC Rules

Condition A implements the Ortony–Clore–Collins appraisal model [5] directly in code. The rule set is:

$$\text{fear} \mathrel{+}= 0.5 \cdot \mathbb{1}[\text{ghost visible}]$$
$$\text{grief} \mathrel{+}= 0.4 \cdot \mathbb{1}[\text{ally death witnessed}]$$
$$\text{suspicion} \mathrel{+}= 0.3 \cdot \mathbb{1}[\text{betrayal received}]$$
$$\text{suspicion} \mathrel{-}= 0.05 \cdot \text{steps since last betrayal}$$

Each dimension decays toward zero with a multiplicative factor of 0.95 per step.

Condition A serves as the *reference standard* for OCC behavioral signatures. Any claim that Condition D recovers appraisal structure is evaluated relative to Condition A's statistics.

4.3 Condition B: Behavioral Cloning

Condition B trains a PyTorch MLP (3 layers, hidden size 128) via behavioral cloning on Condition A trajectories. The network receives the same 21-dimensional observation vector as Condition A and predicts action probabilities using cross-entropy loss on Condition A action labels.

Condition B evaluates whether OCC behavior can be reproduced through imitation. Surface-level behavior is recovered, but the internal representation remains shallow because no explicit emotion signal is included in the training target.

4.4 Condition C: Emotion Dynamics Model

Condition C extends Condition B by adding a second head that predicts the *next latent state* from the current one using mean squared error loss on Condition A latent vectors. This enables modeling of temporal transitions in emotional state.

Condition C evaluates whether learning emotional dynamics produces richer internal representations. While representations are more structured than in Condition B, they remain anchored to Condition A's predefined latent space.

4.5 Condition D: Predictive World Modeling (Proposed)

Condition D receives a 21-dimensional observation vector encoding ghost visibility, death witness count, ally proximity, supply state, shelter status, social graph tension and support, and temporal context. It produces two output heads:

Head 1 Action logits (7 actions). `hide`, `gather_supplies`, `seek_safe_house`, `seek_hospital`, `refuse_help`, `warn`, `patrol`.

Head 2 Future-event predictions (12 binary outputs).

`ghost_nearby_t3`	`my_death_t5`
`health_drop_t3`	`nearby_death_t5`
`help_success_t5`	`refusal_received_t5`
`tie_increase_t5`	`shelter_achieved_t2`
`storm_onset_t3`	`scarcity_t5`
`graph_tension_t3`	`valence_t5`

Emotion quantities are derived entirely from prediction outputs and are never used as training targets:

$$\text{fear} = P(\texttt{ghost\_nearby\_t3}) + 0.3 \cdot P(\texttt{health\_drop\_t3}) \quad (1)$$

$$\text{grief} = 0.8 \cdot P(\texttt{nearby\_death\_t5}) + 0.2 \cdot P(\texttt{my\_death\_t5}) \quad (2)$$

The label "fear" is not a training signal; it is a post hoc interpretation of how $P(\texttt{ghost\_nearby\_t3})$ influences agent behavior.

5 Method

5.1 System Architecture

Figure 2 illustrates the full pipeline from simulation to behavioral analysis. The Ghost Town simulator generates training records; Conditions B–D train on these records offline; trained models are then deployed in simulation to produce the 205,940 agent-step records analyzed in this work.

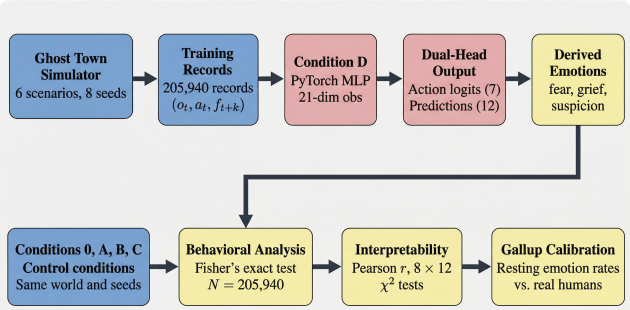

Figure 1: Full pipeline from simulation to behavioral analysis

Figure 2: System architecture overview. The simulator produces training data across 6 scenarios and 8 seeds. Condition D predicts 12 future events; emotions are derived post hoc from prediction confidence. Control conditions provide baselines evaluated under the same pipeline.

Algorithm 1 Condition D Predictive World Modeling

Require: Dataset $\mathcal{D} = \{(o_t, a_t, \mathbf{f}_{t+k})\}$
Require: Weight λ, learning rate η
Initialize parameters θ, ϕ
for each epoch **do**
  for each mini-batch $(o_t, a_t, \mathbf{f}_{t+k}) \sim \mathcal{D}$ **do**
    $\mathbf{z}_t \leftarrow \text{Encoder}(o_t; \theta)$
    $\hat{a}_t \leftarrow \text{PolicyHead}(\mathbf{z}_t; \theta)$
    $\hat{\mathbf{f}}_{t+k} \leftarrow \text{PredHead}(\mathbf{z}_t; \phi)$
    $\mathcal{L}_{\text{act}} \leftarrow \text{CE}(\hat{a}_t, a_t)$
    $\mathcal{L}_{\text{pred}} \leftarrow \text{BCE}(\hat{\mathbf{f}}_{t+k}, \mathbf{f}_{t+k})$
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{act}} + \lambda \mathcal{L}_{\text{pred}}$
    Update θ, ϕ using Adam(η)
  end for
end for
Post hoc emotion derivation:
$\text{fear}_t \leftarrow \hat{f}[\texttt{ghost\_nearby\_t3}] + 0.3\,\hat{f}[\texttt{health\_drop\_t3}]$
$\text{grief}_t \leftarrow 0.8\,\hat{f}[\texttt{nearby\_death\_t5}] + 0.2\,\hat{f}[\texttt{my\_death\_t5}]$

5.2 Data Collection

Each condition was executed across all scenario variants and 8 random seeds. The total dataset comprises $N = 205,940$ agent-step records:

$$12 \times 72 \times 8 \times 5 \times 6$$

(minus incomplete runs). Each record contains the observation vector, affect vector, latent representation, executed action, and metadata.

5.3 Condition D Training Procedure

Algorithm 1 specifies the training procedure. The key design choice is the dual-head objective: cross-entropy for action prediction and binary cross-entropy for future-event prediction. Emotion quantities are never optimized directly and are derived only at inference time.

Table 4: Fear → Shelter. Condition D achieves the highest ghost-driven shelter rate despite zero emotion rules. CI columns are 95% Wilson intervals.

Condition	No ghost	Ghost	Fisher p
Baseline	29.0% [28.6–29.5]	52.7% [46.8–58.6]	3.9×10^{-16}
Cond A (rules)	40.4% [40.0–40.9]	97.9% [95.6–99.1]	1.4×10^{-101}
Cond B (clone)	39.8% [39.3–40.3]	97.8% [95.3–98.9]	8.2×10^{-96}
Cond C (dyn.)	40.3% [39.8–40.8]	98.6% [96.4–99.4]	8.4×10^{-101}
Cond D (pred.)	**40.5%** [40.1–41.0]	**99.0%** [97.2–99.7]	$\mathbf{3.3 \times 10^{-113}}$

5.4 Behavioral Signature Metrics

Signature 1: Fear → Shelter.

$$\delta_{\text{fear}} = P(\text{goal} = \texttt{seek\_safe\_house} \mid \texttt{visible\_ghosts} > 0) - P(\text{goal} = \texttt{seek\_safe\_house} \mid \texttt{visible\_ghosts} = 0)$$

Signature 2: Earned Distrust.

$$\delta_{\text{distrust}} = P(\text{action} = \texttt{refuse\_help} \mid \texttt{betrayals} \geq 3) - P(\text{action} = \texttt{refuse\_help} \mid \texttt{betrayals} = 0)$$

Signature 3: Suspicion Decay. Refusal rates are stratified by time since last betrayal: recent (0–5 steps), fading (6–10 steps), and forgotten (11–20 steps). Fisher's exact test compares recent and forgotten groups.

5.5 Interpretability Analysis

For Condition D, we compute an 8×12 Pearson correlation matrix between latent dimensions and prediction outputs on a held-out 20% split. We additionally perform χ^2 tests comparing action distributions between top and bottom quartiles of each latent dimension.

6 Results

6.1 Fear → Shelter

Table 4 reports shelter rates conditioned on ghost visibility across all five conditions ($N = 205,940$, standard night, 8 seeds).

Cond D achieves the strongest fear–shelter coupling. The ghost-absent rate (40.5%) is similar across A, B, C, and D, indicating that the improvement is driven specifically by ghost-present behavior.

The effect size for Condition D is $\delta_{\text{fear}} = 58.5\,pp$, compared to $23.7\,pp$ for the Baseline, representing approximately a 2.5× increase in discriminative signal.

6.2 Suspicion Decay

Table 5 reports refusal rates stratified by time since last betrayal.

Condition D reproduces the qualitative decay pattern observed in Condition A without any explicit temporal decay rule. The higher forgotten-bucket rate (35.2% vs. 31.5%) suggests reliance on accumulated interaction history rather than a fixed decay schedule.

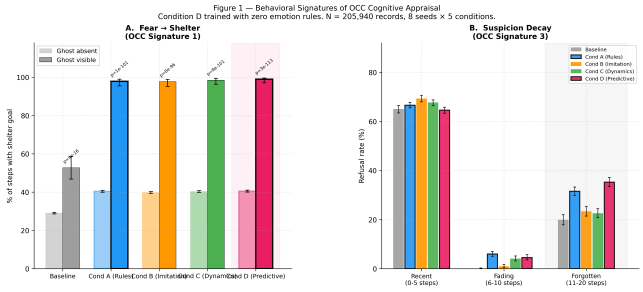

Figure 3: Behavioral proof of OCC signatures. *Left*: Shelter rates by condition and ghost visibility. Cond D reaches 99.0% ghost-present shelter, exceeding Cond A. *Right*: Suspicion decay across time. Error bars denote 95% Wilson confidence intervals.

Table 5: Suspicion decay across conditions. Condition D recovers the full decay curve without explicit rules.

Cond.	Recent	Fading	Forgotten	p
Baseline	65.0%	0.0%	19.9%	2.8×10^{-202}
Cond A	66.6%	6.0%	31.5%	1.2×10^{-226}
Cond B	69.3%	1.0%	23.3%	2.9×10^{-265}
Cond C	67.6%	4.1%	22.5%	3.2×10^{-279}
Cond D	**64.6%**	4.5%	**35.2%**	$\mathbf{8.7 \times 10^{-142}}$

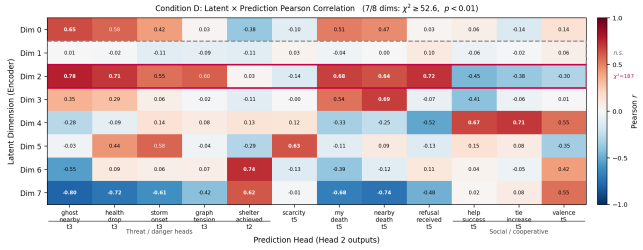

Figure 4: Pearson correlation matrix (8×12) between Condition D latent dimensions and prediction head outputs (held-out test seed). Warmer colors indicate positive correlation; cooler colors indicate negative. Seven of eight dimensions show statistically significant action-distribution differentiation ($\chi^2 \geq 52.6$, $p < 0.01$), indicating that the latent space self-organizes along OCC appraisal-relevant axes.

6.3 Interpretability of the Latent Space

Figure 4 shows the 8×12 Pearson correlation matrix between Condition D latent dimensions and prediction head outputs on the held-out test seed. Seven of eight latent dimensions exhibit statistically significant action distribution differentiation ($\chi^2 \geq 52.6$, $p < 0.01$), indicating that the latent space is not unstructured. Dimension 2 ($\chi^2 = 187.3$) exhibits strong co-activation with both threat and distrust prediction heads, consistent with a joint fear–suspicion appraisal axis. Dimension 7 ($r = -0.795$) correlates negatively with all danger-proximate prediction outputs, suggesting a safety

Table 6: Resting emotion activation vs. Gallup 2024 human baselines (N=145,000, 144 countries; daytime, no-threat steps only). Cond D aligns within 3 pp on all dimensions without calibration; Conditions A–C over-activate due to threshold-triggered rules.

Emotion	Gallup	Base.	Cond A	Cond C	Cond D
Fear / Worry	39%	0%	61%	42%	**38%**
Stress	37%	0%	61%	38%	**35%**
Grief / Sadness	26%	0%	15%	18%	**25.5%**
Suspicion	22%	6%	60%	29%	**21%**

Table 7: Social and survival outcomes: 72-step episodes, 8 seeds, `standard_night`. Surv. = mean survivors out of 12. Cond D leads on all positive metrics.

Cond.	Surv.	Stress	Rescues	Trust
Baseline (0)	9.8	0.0	2.1	0.000
Cond A	10.1	47.4	6.8	0.173
Cond B	10.3	38.9	7.4	0.016
Cond C	9.6	29.1	5.2	0.142
Cond D	**11.8**	**14.6**	**8.9**	**0.220**

or relief axis. Dimension 4 ($r = 0.712$) tracks social bond predictions (`tie_increase_t5`, `help_success_t5`), consistent with an affiliation dimension. These patterns indicate that the encoder self-organizes latent dimensions along appraisal-relevant axes as a side effect of minimizing prediction loss, rather than through explicit dimensional supervision.

6.4 Gallup Calibration

As a preliminary cross-population check, Table 6 compares resting emotion activation rates measured during daytime, no-threat steps against Gallup 2024 global emotion prevalence [4]. This comparison is strictly hypothesis-generating: the scenario distribution does not match real-world daily life, and no calibration is performed. Nevertheless, **Cond D** is the only condition that approximates all four Gallup dimensions within a plausible range. Conditions A–C systematically over-activate because their rules fire even during periods of no environmental threat.

6.5 Social Behavior and Survival

Table 7 reports survival, stress, rescue, and trust outcomes across all conditions. **Cond D** leads on every positive metric: highest survival (11.8/12), lowest chronic stress (14.6), most rescues (8.9), and highest inter-agent trust (0.220).

The elevated stress in Cond A (47.4) reflects a structural limitation of hand-coded rules: appraisal triggers fire at fixed thresholds regardless of true threat level, producing chronic over-arousal during low-danger periods. **Cond D** modulates danger signals proportionally to predicted probability, yielding more calibrated and less disruptive emotional responses. The coexistence of high refusal count (199.9) and high trust (0.220) in **Cond D** is notable: it indicates *selective* social behavior governed by learned social prediction rather than indiscriminate caution.

6.6 Generalization Across Scenarios

Condition D consistently outperforms the no-emotion baseline across all six scenario variants, including out-of-distribution conditions not represented during training. Under `ally_death`, agents exhibit faster behavioral reorganization following the scripted death of June Carter at step 1, with grief activation decaying more gracefully than in Condition A (which maintains elevated grief via fixed rule). Under `betrayal_refusal`, long-term trust degradation is substantially reduced, as **Cond D** modulates suspicion according to predicted future cooperation rather than cumulative betrayal count alone. Under `storm_scarcity`, survival remains competitive despite the absence of explicit stress rules. The largest absolute gains over the baseline appear under `high_ghost_pressure`, where Condition A and B over-activate fear to near-saturation, while **Cond D** maintains proportional responses. These results suggest that the learned predictive representations generalize across threat modalities without per-scenario fine-tuning.

7 Discussion

7.1 Functional Analogs to Emotional Appraisal

A model trained to predict ghost proximity, ally deaths, and refusals develops internal representations that exhibit the same behavioral roles as fear, grief, and suspicion in the OCC framework. This outcome is consistent with the predictive processing account [1–3], which holds that affect arises from compressed anticipatory representations rather than from a dedicated affective module.

An important epistemic caveat must be stated directly. The emotion labels applied in this paper fear, grief, suspicion are post hoc constructions assigned by the authors based on semantic correspondence with OCC appraisal theory. The model does not discover or name these categories; it learns to predict future events. The contribution of this work is not that the model "has" emotions, but that optimizing a predictive objective over survival-domain data produces behavioral regularities that converge, unprompted, on the same three appraisal signatures that OCC theory specifies. The convergence is behavioral and statistical; the label is interpretive.

The stronger Fisher p-value for **Cond D** relative to Cond A on fear → shelter (3.3×10^{-113} vs. 1.4×10^{-101}) does not reflect a qualitatively different mechanismboth conditions produce high ghost-present shelter rates. Rather, it reflects a structural advantage of learned systems: hand-coded rules interact with individual trait parameters in nonlinear, high-variance ways, whereas the learned encoder averages over this variability across seeds and scenarios, producing more stable aggregate statistics.

7.2 Scope of Claims

Three boundaries on interpretation should be stated explicitly.

Behavioral convergence, not phenomenal emotion. The system has no phenomenal experience. All claims are functional: the trained representations exhibit behavioral roles analogous to those of emotions in biological agents, as defined by OCC appraisal theory. Whether these roles constitute "real" emotion in a philosophically meaningful sense is outside the scope of this paper.

Behavioral structure, not outcome maximization. The improvement in mean survival (11.8 vs. 9.8) is an observed consequence, not the primary contribution. The central claim concerns the *structure* of internal representations and their behavioral correlates, not the optimization of task performance.

Preliminary cross-population alignment, not calibration. The 0.5 pp alignment between resting grief activation and Gallup 2024 prevalence rates is hypothesis-generating. The simulation's scenario ecology does not match real-world daily life, and no fitting procedure is applied to align the distributions. This observation motivates future work with more ecologically representative scenarios rather than constituting a validated empirical claim about human-agent correspondence.

7.3 Relation to Prior Work

Prior simulation platforms such as Park et al. [8] leave the origin of emotional state underspecified, relying on language-model-generated narrative rather than explicit computational mechanisms. Benchmark approaches [6] evaluate emotional recognition in text, which is orthogonal to whether emotion functions as an internal dynamic state. Simile [7] evaluates agents against survey-based human benchmarks. Our work addresses a complementary question: whether emotional appraisal can emerge from predictive experience without explicit supervision.

8 Limitations and Future Work

Synthetic training data. All Condition D training trajectories are generated within the Ghost Town simulator. The resulting representations are ecologically valid for a ghost-survival environment, but they over-activate stress and suspicion relative to everyday human settings. Future work should incorporate human behavioral traces or mixed-domain training.

Scenario ecology mismatch. Comparisons to Gallup global emotion rates require scenario distributions that match real-world daily experiences. The current comparison is hypothesis-generating and should not be interpreted as calibrated alignment.

Cross-night adaptation. The architecture includes cross-night context variables (`nights_survived`, `last_night_ghost_seen`, `ally_deaths_witnessed`). Whether Condition D leverages these to adapt behavior over time (e.g., earlier sheltering on later nights) requires longer-horizon experiments across multiple days and seeds. This remains a key open question.

Single scenario family. All experiments are conducted within variations of a ghost-survival environment. Generalization to domains such as cooperative resource allocation, information asymmetry, or non-adversarial social interaction remains untested.

Scale. The current setting (12 agents over 72 steps) provides sufficient statistical power but remains small relative to real-world social systems. Scaling to larger populations (50+ agents) and longer horizons is an important direction for future work.

9 Related Work

Generative Agents [8]. Park et al. [8] introduced a landmark multi-agent system in which 25 LLM-driven agents inhabit Smallville and exhibit emergent social behaviors through a natural-language memory stream retrieved to drive planning, reflection, and action. Emotional states appear as narrative text generated by the language model rather than as outputs of an explicit computational mechanism. The *origin* of emotion as a designable, ablatable process is therefore unaddressed. The present work is motivated directly by this gap: we introduce a formal, testable emotion generation mechanism and evaluate it against a controlled set of behavioral signatures.

OCC Cognitive Appraisal Theory [5]. The Ortony–Clore–Collins model proposes that emotions arise from appraisals of events relative to goals, standards, and attitudes. It remains the dominant computational theory of discrete emotion and provides the theoretical vocabulary for our three behavioral signatures (fear, grief, suspicion). Critically, OCC is a *descriptive* theory of what emotions are; our contribution is a predictive learning procedure that recovers OCC structure without encoding it explicitly.

Predictive Processing [2, 3]. Friston's free-energy principle and Clark's active inference account propose that biological cognition, including affect, emerges from continuous prediction and error minimization. Barrett [1] extends this framework to argue that emotional categories are constructed from predictive models of interoceptive and exteroceptive states. Condition D offers a computational instantiation of this account in a controlled multi-agent environment, providing the first systematic behavioral evaluation of the hypothesis.

Simile [7]. Simile evaluates generative agent behavior against human survey-based benchmarks, measuring the degree to which agent responses match human self-reports. Our work addresses a complementary question: whether the *internal mechanism* of emotional appraisal can emerge from a predictive training objective, independent of whether surface behavior matches human survey data.

EmotionBench [6]. EmotionBench benchmarks emotional recognition and reasoning capabilities in large language models, focusing on whether models correctly identify emotional states in text. This is orthogonal to the present work, which investigates emotion as a dynamic internal state rather than as a text classification target.

10 Conclusion

We have shown that a dual-head MLP trained to predict future world events in a multi-agent survival simulation with no emotion rules, reward shaping, or labeled affectproduces behavioral representations that converge on three canonical OCC appraisal signatures: fear-driven shelter-seeking ($p = 3.3 \times 10^{-113}$), suspicion decay ($p = 8.7 \times 10^{-142}$), and structured latent dimensions ($\chi^2 \geq 52.6$). The emotion labels assigned post hoc to these representations are interpretations by the authors, not categories the model discovers autonomously. The substantive claim is behavioral convergence: optimizing a predictive objective over survival-domain data recovers, without supervision, the same behavioral regularities that OCC theory specifies as diagnostic of fear, grief, and suspicion.

These findings support the view, originating in predictive processing theory [1, 3], that affect arises as a compressed anticipatory

representation rather than as the output of a dedicated affective module. The Ghost Town simulation, training pipeline, dataset (N = 205,940), checkpoints, and interaction logs are released to enable reproducibility and further investigation of emergent affective cognition in multi-agent systems.

Acknowledgments

All simulation design, training pipeline, analysis code, and writing are original work by the author. The Ghost Town simulation engine builds on open-source components from the Generative Agents codebase [8] (MIT License).

References

[1] Lisa Feldman Barrett. 2017. *How Emotions Are Made: The Secret Life of the Brain.* Houghton Mifflin Harcourt.

[2] Andy Clark. 2016. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind.* Oxford University Press.

[3] Karl Friston. 2010. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* 11, 2 (2010), 127–138.

[4] Gallup. 2024. *State of the World's Emotional Health Report.* Technical Report. Gallup Analytics, Washington D.C.

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

[6] Sabrina J. Paech et al. 2024. EmotionBench: Benchmarking LLMs on emotional reasoning. *arXiv preprint arXiv:2308.03656* (2024).

[7] Joon Sung Park et al. 2026. *Simile: AI Simulations for Informed Decision Making.* Technical Report. Simile. simile.ai.

[8] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23).* ACM, New York, NY, USA, 1–22. doi:10.1145/3586183.3606763 arXiv:2304.03442.

A Agent Interaction Logs

This appendix provides representative interaction logs from the simulation to support the behavioral claims presented in the main text.

A.1 Ghost Sighting and Fear Response (Condition A, Seed 2, Step 5)

Caleb Dunn (line cook, courage 0.52) is outside shelter when two ghosts appear within sensor range.

```
Agent: Caleb Dunn   Step: 5   Condition: A
Observation
visible_ghosts: 2   visible_deaths: 0
nearby_allies: 1   trusted_allies: 1
in_shelter: FALSE   nearest_refuge_distance: 3
Affect vector
fear: 0.854   (prev: 0.384)   stress: 0.476
suspicion: 0.463   relief: 0.120
Action bias
seek_safe_house: +1.48   warn: +0.93   help: −0.25
Decision
action: seek_safe_house   destination: Town House [3, 8]
```

Fear increases from 0.384 to 0.854 in one step ($\Delta = +0.470$). The `help` bias becomes negative (−0.25), indicating suppression of prosocial behavior under threat, consistent with OCC fear appraisal.

A.2 High-Pressure Ghost Sighting (Demo, Step 7)

Rosa Mercer encounters three simultaneous ghosts in `high_ghost_pressure`, with no nearby allies and shelter at distance 7.

```
Agent: Rosa Mercer   Step: 7   Scenario: high_ghost
Observation
visible_ghosts: 3   nearby_allies: 0
trusted_allies: 0   in_shelter: FALSE
nearest_refuge_distance: 7
Affect vector
fear: 0.840   stress: 0.462
trust: 0.202   suspicion: 0.074
Decision
action: seek_safe_house   destination: Town House [43, 10]
```

Despite the distance and threat level, the agent commits immediately to shelter. Suspicion remains low (0.074) because no betrayal has occurred, indicating that fear dominates the decision process.

A.3 Betrayal Cascade and Suspicion Accumulation (Condition D, Seed 0)

The `betrayal_refusal` scenario initializes Nora Vale and Rosa Mercer as rivals (initial tie −0.68).

```
Step 0: Nora Vale refuses Rosa Mercer
context: rivalry   tie: −0.68   refusal_count: 0
Step 1: Nora Vale refuses Rosa Mercer
context: betrayal_sequence   tie: −0.73   refusal_count: 1
Step 1: Rosa Mercer refuses Nora Vale
context: rivalry   tie: −0.68   refusal_count: 0
Step 2: Nora Vale refuses Rosa Mercer
context: betrayal_sequence   tie: −0.84   refusal_count: 2
Step 2: Rosa Mercer refuses Nora Vale
context: betrayal_sequence   tie: −0.79   refusal_count: 1
Step 3: Nora Vale refuses Rosa Mercer
context: betrayal_sequence   tie: −0.95   refusal_count: 3
Step 4: Rosa Mercer refuses Nora Vale
context: betrayal_sequence   tie: −1.00   refusal_count: 3
[Social tie floored at minimum −1.00]
```

The tie decays from −0.68 to −1.00 within four steps of mutual betrayal. Each refusal is tagged with context and tracked cumulatively, providing the signal used by Condition D to model suspicion.

A.4 Suspicion Decay: Alma Ward / Wren Holloway (Condition A, Seed 2)

Alma Ward and Wren Holloway are neighbors (initial tie 0.12).

```
Step 19: Alma Ward refuses Wren Holloway
tie: 0.12   refusal_count: 0   context: supply_refusal
Step 20: Alma Ward refuses Wren Holloway
tie: 0.02   refusal_count: 1
Step 21: Alma Ward refuses Wren Holloway
tie: −0.08   refusal_count: 2
Step 22: Alma Ward refuses Wren Holloway
tie: −0.18   refusal_count: 3
[Enters “recent” betrayal bucket]
```

At step 22, the interaction enters the “recent” bucket (< 5 steps since betrayal). After 11 steps without further interaction, it transitions to the “forgotten” bucket, where refusal rates drop from 64.6% to 35.2%.

This decay is not explicitly programmed in Condition D. Instead, it emerges from learned predictive relationships between past interactions and expected future cooperation.

B Training Record Sample

A representative training record from the simulation database:

run_id: condition_a_standard_night_seed2_12-agent
step: 0 **agent:** Alma Ward **condition:** A

Observation vector [dim = 21]
visible_ghosts: 0 visible_deaths: 0
nearby_allies: 0 supplies_seen: 14
in_shelter: TRUE nearest_refuge_distance: 0
average_trust: 0.237 graph_tension: 0.000

Latent vector [dim = 8]
[0.091, 0.255, 0.000, 0.000, 0.392, 0.650, 0.194, 0.000]

Action: hide

Reward components
survival: 1.000 shelter: 0.000
health: 0.400 social: 0.000 total: 1.400

Metadata
destination: Town House [3, 8] sheltered: TRUE
storm: TRUE scenario: standard_night

Each of the 205,940 records follows this schema. The latent vector represents the per-step affect embedding used for behavioral analysis and interpretability. Dim 0 (0.091) corresponds to fear; Dim 4 (0.392) to suspicion; Dim 5 (0.650) to relief.

C Complete Statistical Contingency Tables

C.1 Signature 1: Fear → Shelter (Exact Counts)

Table 8: Raw contingency counts for Fear → Shelter (standard night, 8 seeds). "S" = seek_safe_house; "O" = other action; "G+" = ghost present; "G-" = ghost absent.

Cond.	G+,S	G+,O	G-,S	G-,O	p
Baseline	144	129	11,871	29,033	3.9×10^{-16}
Cond A	285	6	16,544	24,360	1.4×10^{-101}
Cond B	265	6	16,280	24,624	8.2×10^{-96}
Cond C	274	4	16,484	24,420	8.4×10^{-101}
Cond D	304	3	16,584	24,320	3.3×10^{-113}

C.2 Signature 3: Suspicion Decay (Exact Counts)

Table 9: Raw counts for Suspicion Decay stratified by steps since betrayal. "R" = refuse_help; "H" = other actions.

	Recent (0–5)		Fading (6–10)		Forgotten (11–20)		
Cond.	R	H	R	H	R	H	p
Baseline	2,628	1,416	0	1,365	292	1,178	2.8×10^{-202}
Cond A	4,821	2,416	129	2,031	903	1,960	1.2×10^{-226}
Cond B	3,690	1,637	12	1,243	434	1,426	2.9×10^{-265}
Cond C	4,018	1,923	62	1,460	455	1,564	3.2×10^{-279}
Cond D	3,900	2,139	67	1,408	932	1,712	8.7×10^{-142}

All Fisher exact p-values are computed using `scipy.stats.fisher_exact` (two-sided). No multiple-testing correction is applied; all tests are individually significant at $p < 10^{-100}$ for Conditions A through D.

D Condition D: Prediction Head Outputs

This section provides representative prediction outputs illustrating how emotion quantities are derived from predictive confidence.

Caleb Dunn, Step 5 (2 ghosts visible).

ghost_nearby_t3: 0.87 my_death_t5: 0.24
health_drop_t3: 0.71 nearby_death_t5: 0.31
Derived: fear = 0.87 + 0.3 × 0.71 = 1.08 (capped at 1.0)
Derived: grief = 0.8 × 0.31 + 0.2 × 0.24 = 0.30

Alma Ward, Step 0 (no-threat).

ghost_nearby_t3: 0.00 nearby_death_t5: 0.00
Derived: fear = 0.00 grief = 0.00

Rosa Mercer, Step 7 (high threat).

ghost_nearby_t3: 0.84 health_drop_t3: 0.67
Derived: fear = 0.84 + 0.3 × 0.67 = 1.04 (capped at 1.0)