Conceptual Model: Adaptive Memory in Norm Formation

1 Research Motivation

Social norms emerge from repeated interactions, but individuals have limited memory. Existing models typically assume fixed memory capacity independent of interaction outcomes. However, in reality, successful coordination increases confidence and reliance on past experience, while failure triggers uncertainty and discounting of old information.

Core Question: How does outcome-dependent memory adaptation affect the speed, stability, and flexibility of norm emergence?

2 Core Concept

Memory is not passive storage—it is actively modulated by prediction accuracy.

Traditional:	Memory → Belief → Decision	(fixed capacity)
Our Model:	Memory ↔ Trust ↔ Prediction Accuracy	(adaptive capacity)

When an agent interacts with a partner:

1. **Predict** partner's action (based on memory)
2. **Observe** actual action
3. Compute **prediction error**
4. Update **confidence/anxiety** level
5. Confidence determines **how much history to consider**

3 Agent Mental State

Variable	**Notation**	**Description**
Memory	$M(t)$	Record of past interactions
Belief	$\mathbf{b} = [P(A), P(B)]$	Estimated strategy distribution
Temperature	$\tau \in [\tau_{min}, \tau_{max}]$	Anxiety/uncertainty level
Trust	trust $= f(\tau)$	Confidence in environment stability

Temperature interpretation:

- Low τ: Confident, consistent behavior (exploitation)
- High τ: Anxious, exploratory behavior (exploration)

Trust-Temperature relationship:

$$\text{trust} = 1 - \frac{\tau - \tau_{min}}{\tau_{max} - \tau_{min}} \tag{1}$$

4 The Dual Feedback Loops

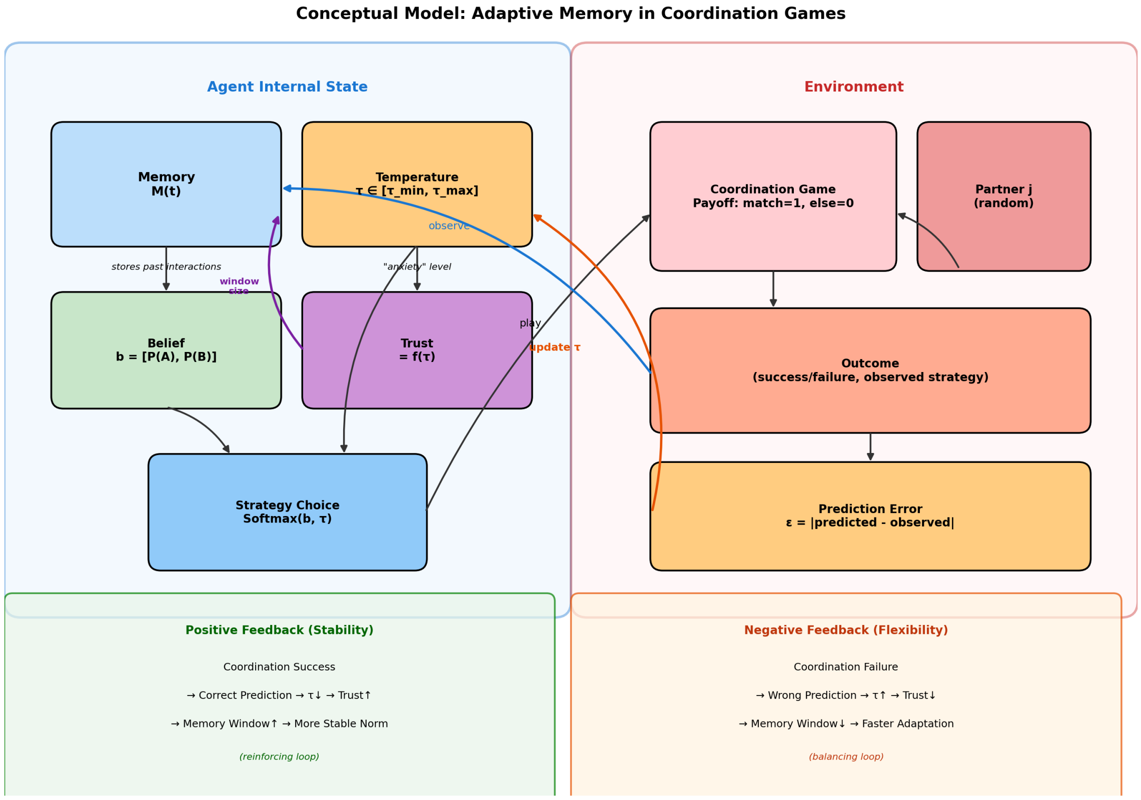

Figure 1: Conceptual model showing agent internal state, environment interaction, and dual feedback loops. The positive loop (green) reinforces stability through successful coordination; the negative loop (orange) enables flexibility through failure-triggered adaptation.

4.1 Positive Feedback (Reinforcing Loop)

$$\text{Coordination Success} \rightarrow \text{Correct Prediction} \rightarrow \tau \downarrow \rightarrow \text{Trust}\uparrow \rightarrow \text{Longer Memory} \rightarrow \text{Stable Beliefs} \rightarrow \text{More Success}$$

Effect: Success breeds success. Once coordination begins, it self-reinforces.

4.2 Negative Feedback (Balancing Loop)

$$\text{Coordination Failure} \rightarrow \text{Wrong Prediction} \rightarrow \tau \uparrow \rightarrow \text{Trust}\downarrow \rightarrow \text{Shorter Memory} \rightarrow \text{Adaptive Beliefs} \rightarrow \text{Exploration}$$

Effect: Failure triggers adaptation. The system can escape suboptimal equilibria.

5 Memory Types

Dynamic Memory (our innovation):

$$\text{window} = \text{base} + \lfloor \text{trust} \times (\text{max} - \text{base}) \rfloor \tag{2}$$

Upper bound (≈ 6) reflects human working memory limits (Miller's Law).

Type	Window	Weights	Trust Effect
Fixed	Constant k	Equal: $w = 1/k$	None
Decay	Soft (effective)	Exponential: $w = \lambda^{age}$	None
Dynamic	Variable $[base, max]$	Equal over window	Window adapts with trust

6 Decision and Learning

6.1 Action Selection

Agent chooses action via temperature-modulated softmax:

$$P(\text{choose } A) = \frac{\exp(b_A/\tau)}{\exp(b_A/\tau) + \exp(b_B/\tau)} \tag{3}$$

6.2 Temperature Update

$$\tau_{t+1} = \begin{cases} \max(\tau_{min}, \tau_t \times (1-\alpha)) & \text{if prediction correct (cooling)} \\ \min(\tau_{max}, \tau_t + \beta) & \text{if prediction wrong (heating)} \end{cases} \tag{4}$$

Asymmetry: Confidence builds gradually (multiplicative), breaks quickly (additive). This reflects psychological findings on trust dynamics.

7 Theoretical Predictions

1. **Faster Convergence**: Dynamic memory should accelerate norm emergence due to stronger positive feedback.

2. **Greater Resilience**: Established norms under dynamic memory should resist transient perturbations (large window creates inertia).

3. **Better Adaptation**: When environment changes, dynamic memory enables faster re-equilibration (failure shrinks window, accelerating belief updates).

8 Relationship to Literature

Literature	Connection
Bounded Rationality (Simon)	Finite memory, satisficing via softmax
Reinforcement Learning	Prediction error drives updates
Cultural Evolution (Young, Boyd)	Norm emergence through adaptive play
Trust Dynamics	Slow build, fast destruction pattern

Our contribution: Memory as an *endogenous*, outcome-dependent cognitive process, not merely an exogenous constraint.

9 Discussion Questions

1. Is the trust-memory link psychologically plausible? What empirical evidence exists?

2. What determines optimal feedback loop strengths (cooling rate vs. heating penalty)?

3. Under what conditions would fixed memory outperform dynamic memory?

4. How does network structure affect these dynamics beyond mean-field?
5. Normative implications: Is faster convergence always desirable?

Summary

Agents form beliefs from memory, make decisions modulated by anxiety (τ), and update anxiety based on prediction errors. Anxiety determines trust, which modulates memory window size in dynamic memory. This creates dual feedback loops: **success reinforces stability**, **failure enables flexibility**. The interplay determines norm emergence dynamics.