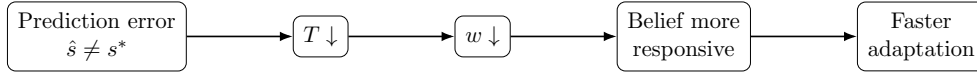


Concept Diagrams: Two Feedback-Loop Variants

Variant 2: Purely Cognitive Loop (Window-only)

Negative/balancing



Positive/reinforcing

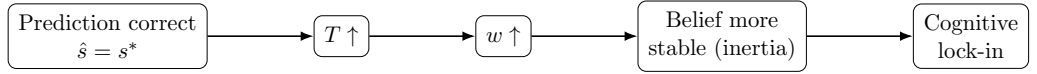
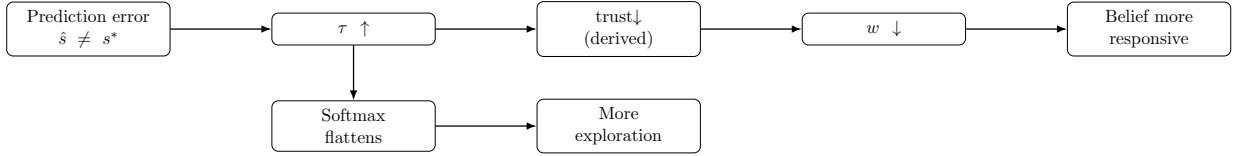


Figure 1: Variant 2 (draft2): Trust T only modulates memory window size w , which changes how quickly beliefs respond. There is no separate decision-noise channel.

Variant 1: Cognitive + Decision Channel (Window + Choice Concentration)

Negative / balancing loop



Positive / reinforcing loop

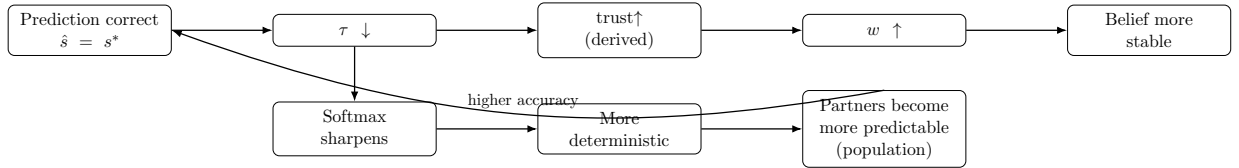


Figure 2: Variant 1 (main): τ affects both (i) memory window w via derived trust and (ii) decision noise via softmax. The behavioral channel (softmax sharp/flat) is shown separately from the memory-window channel to avoid overlap.

Internal Memo (not for sharing)

Big-picture motivation

This project studies the cognitive micro-foundations of norm / institution emergence in repeated random matching. The core idea is that agents adapt the *timescale* of belief formation (their effective memory window) based on how predictable the environment seems.

A robust behavioral regularity is that when environments are volatile or uncertain, people rely more on recent information (shorter integration window / higher learning rate); when environments are stable, they integrate over longer horizons (longer window / lower learning rate). We model this with an endogenous window size $w(t)$.

Why a minimal mechanism (vs. active inference)

Active inference can in principle capture similar phenomena via precision/uncertainty and policy selection, but it introduces additional modeling commitments (generative model structure, priors over policies, utility/priors over outcomes). Here the goal is intentionally narrower: isolate a minimal feedback loop that can generate (i) stabilization/lock-in of beliefs and (ii) flexibility under prediction failure.

Accordingly, we use prediction correctness as a lightweight error signal and ask: *is uncertainty-driven adjustment of memory timescale sufficient to produce norm stabilization?* If yes, more elaborate Bayesian/active-inference formulations can be added later as extensions.

Two variants and what changes

Variant 2 (window-only, “cognition-only”). Internal state is trust $T \in [0, 1]$. Prediction errors decrease T , correct predictions increase T . Trust controls only the effective memory window $w(T)$, which changes how quickly beliefs respond. There is no separate behavioral amplification beyond what beliefs imply.

Variant 1 (window + decision channel). Internal state is decision temperature τ . Prediction errors increase τ (more randomness/exploration), correct predictions decrease τ (more deterministic choice). A derived trust variable (monotone transform of τ) sets the memory window. This adds a behavioral channel that can reinforce predictability at the population level, often speeding convergence.

How to discuss with an advisor (one-minute version)

- Both models: prediction error \rightarrow internal uncertainty/trust \rightarrow memory window $w \rightarrow$ belief stability vs. responsiveness.
- Window-only: stabilization comes from belief inertia (cognitive lock-in).
- Window+decision: adds softmax concentration as a behavioral channel, strengthening positive feedback and accelerating convergence.
- Use window-only as a clean baseline / ablation; use window+decision as the behaviorally richer main model.