Correction Freeze Problem

Undetected Drift, Hidden Capture, and Limits of Correction

Quan-Hoang Vuong[1,2,3]

[1]The Centre for Interdisciplinary Social Research, Phenikaa University, Hanoi, Vietnam
[2]The University College, Korea University, Seoul, South Korea
[3]Centre Emile Bernheim de Recherche Interdisciplinaire en Gestion, Université Libre de Bruxelles, Brussels, Belgium

September 9, 2026

Abstract

This paper proves four independent conditions under which a system's perceived state can become detached from its true state, and shows what follows when that detachment coincides with self-reinforcing dynamics that make an eventual correction harder to reverse than it was to prevent. A sequence of individually imperceptible changes can accumulate to arbitrary cumulative drift without any single step ever being flagged. Two independent, fully rational reasons can lead a corrective judgment to be legitimately suspended: genuine uncertainty about the reliability of one's own instrument, and a collapse of verification incentive below a sharp cost threshold. Even a fully active, fully resourced, fully self-trusting corrector faces a hard information-theoretic ceiling on accuracy once the signal distinguishing genuine change from ordinary variation grows statistically indistinguishable. A population state governed by a bandwagon-reinforced dynamic exhibits a trichotomy of long-run outcomes separated by a hysteresis gap, so reversing an already-shifted state requires strictly more than what triggered the shift. Together, these results suggest a precise reading of familiar worries about smartphone habits, synthetic media, and self-reinforcing online communities, not as vague metaphors, but as specific, checkable conditions under which an unnoticed and hard-to-reverse drift becomes possible.

Keywords: baseline drift; verification cost; ambiguity aversion; hysteresis; bifurcation theory; correction dynamics

1 Introduction: The Correction Freeze Problem

Can a system's belief about its own state become detached from what is actually true, and stay detached, not from negligence but as the necessary consequence of how correction itself works? Call this the correction freeze problem. A corrective mechanism compares what it currently believes, a perceived state, against what is currently true, and updates only when the two diverge enough to notice. The question this paper answers is when that comparison can fail systematically, not through error, but through one of several structurally distinct, individually rational mechanisms, so that the true state moves without bound while the perceived state stands still, or updates in a way that carries no real information about where the truth has gone.

This is not only a mathematical curiosity. A habit, a norm, or a belief can shift through increments too small to notice, or against a backdrop too costly to keep checking, or through content too statistically similar

to genuine content to reliably tell apart, and any of these can combine with dynamics that make the shift progressively harder to reverse the longer it goes unnoticed. Section 5 reads the framework against smartphone use, synthetic media, and information overload; the point here is only that the underlying question, whether and how correction can freeze, is not an idle one.

Section 2 sets up the shared formal object, a true state and a perceived state connected by an update rule. Section 3 proves four logically distinct routes by which that rule's correction branch fails to fire. Section 4 shows what happens when a frozen perception meets a state whose own dynamics reinforce themselves: an unnoticed drift can complete an essentially irreversible shift, a hidden capture, before it is ever perceived to have moved. Section 5 discusses what the results might illuminate about real cases, and where they stop. Section 6 concludes.

2 Formal Setup

Fix a discrete time index $n = 0, 1, 2, \ldots$ Let $x_n \in [0, 1]$ denote the true state of a system, and $\hat{x}_n \in [0, 1]$ the perceived state, the value a corrective mechanism currently believes x_n to be. The two coincide at time 0. At each step the true state receives an increment $g_n := x_n - x_{n-1}$, taken to be nonnegative and same-signed for simplicity.

Definition 1 (Update rule)**.** *$\hat{x}_n = x_n$ if the mechanism is active at step n and $|x_n - x_{n-1}| \geq \tau$; otherwise $\hat{x}_n = \hat{x}_{n-1}$, where $\tau > 0$ is a fixed perceptibility threshold and active means the mechanism is actually attempting the comparison. The trigger compares consecutive true states, not the true state against the current belief: this is what lets a sequence of small true steps go unflagged indefinitely (Section 3.1) while still allowing $\hat{x}_n$ to track x_n exactly whenever a step is large enough to notice.*

Section 3 shows this rule's correction branch can fail to fire in four distinct ways, three producing an *exact* freeze, $\hat{x}_n$ never updates, and a fourth producing a *statistical* freeze, updates occur but carry no reliable information.

3 Four Routes to Correction Freeze

There are four logically distinct routes by which the update rule's correction branch can fail to fire: a **threshold** route, the drift is too small to trigger correction; a **cost** route, correction is not worth doing; an **ambiguity** route, correction is rationally substituted away; and an **information limit** route, correction cannot reliably distinguish drift from noise even when fully attempted. Each is proved below from independent first principles; none presupposes any of the others.

3.1 Threshold

Proposition 2 (Shifting-Baseline Ratchet, SBR)**.** *For any target cumulative distortion $D \in (0, 1 - x_0]$ and any perceptibility threshold $\tau > 0$, there exists a finite N and a sequence of increments $g_1, \ldots, g_N$, each satisfying $g_n < \tau$, such that no single increment ever clears the threshold, yet the cumulative effect is exactly D: $x_N - x_0 = D = \sum_{n=1}^{N} g_n$.*

Proof. Fix $D \in (0, 1 - x_0]$, $\tau > 0$, and choose $N > D/\tau$. Set $g_n = D/N < \tau$ for every n, all same-signed. Then $x_N - x_0 = \sum_{n=1}^{N} g_n = D$, a telescoping sum in which every intermediate term cancels, and $x_N = x_0 + D \leq 1$ stays in the domain. The bound $D \leq 1 - x_0$ is a bookkeeping constraint carried over from the state space $[0, 1]$ used to connect this result to Section 4, not a limitation of the mechanism itself: run in $x_n \in \mathbb{R}_{\geq 0}$ instead, the identical telescoping argument permits any $D > 0$ whatsoever, however small τ is fixed to be. No finite ceiling on cumulative drift ever follows from τ alone. □

Corollary 3 (Freeze under SBR)**.** *If the increments of Proposition 2 drive x_n and the mechanism is active at every step, then $\hat{x}_n \equiv \hat{x}_0$ for all $n \leq N$, since $|x_n - x_{n-1}| = g_n < \tau$ at every step: the trigger compares consecutive true states, so it never sees more than one increment at a time, however far x_n has drifted from x_0.*

The increment g_n may combine several distinct sources, authored, endogenous, or both, without affecting the argument, which only uses that each step is small and same-signed.

3.2 Cost

Assumption 4. *Let $k \geq 0$ be verification effort, $c > 0$ its constant marginal cost, and $\Lambda(k) = \Lambda_0 e^{-\rho k}$ the expected residual loss after k units of checking. Let $g(L) > 0$ be strictly increasing in a visible-range variable $L > 0$ (how wide the observable range of ordinary variation looks). Total cost is $C(k) = ck + \Lambda_0 g(L) e^{-\rho k}$.*

Theorem 5 (Verification collapse)**.** *Under Assumption 4, the cost-minimizing verification level is $k^*(L) = \max\{0,\ \rho^{-1} \ln(\rho \Lambda_0 g(L)/c)\}$, and there is a threshold $L^\dagger$, solving $\rho \Lambda_0 g(L^\dagger) = c$, below which $k^*(L) = 0$ exactly.*

Proof. C is strictly convex in k, so the unconstrained minimizer $k = \rho^{-1} \ln(\rho \Lambda_0 g(L)/c)$ is optimal whenever nonnegative; otherwise C is already increasing at $k = 0$, so the constrained minimum is the corner $k^* = 0$, by the standard corner-solution argument for constrained convex minimization (see e.g. Boyd and Vandenberghe, 2004). More generally, for any strictly convex, strictly decreasing loss Λ with finite $\Lambda'(0)$, the same argument applied to the general first-order condition $c = -\Lambda'(k) g(L)$ shows the corner solution holds exactly below the threshold $L^\dagger$ solving $c = -\Lambda'(0) g(L^\dagger)$; specializing to $\Lambda(k) = \Lambda_0 e^{-\rho k}$ gives $-\Lambda'(0) = \rho \Lambda_0$ and recovers $\rho \Lambda_0 g(L^\dagger) = c$ as stated above. □

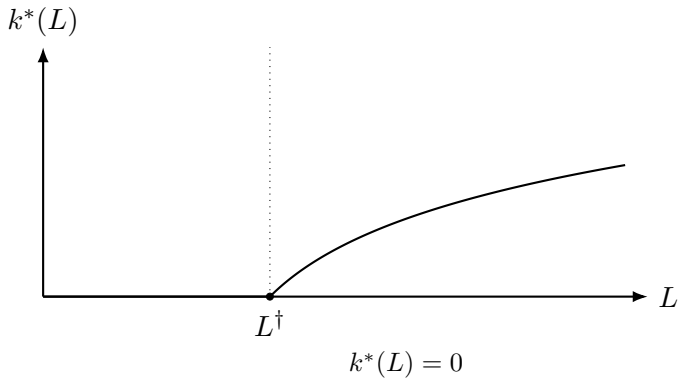

Figure 1: Verification effort $k^*(L)$ against the visible-range variable L (Theorem 5; curve shape illustrative, since $g(L)$ is left general). Below $L^\dagger$, $k^* = 0$ exactly, not merely small: the transition at $L^\dagger$ is a corner, not a gradual thinning. Above $L^\dagger$, effort rises smoothly as the perceived stakes grow.

Corollary 6 (Freeze under verification collapse)**.** *If $L < L^\dagger$ at every step, verification effort is exactly zero throughout, so $\hat{x}_n \equiv \hat{x}_0$ for all n, regardless of the size of the increments g_n.*

A narrow visible range is one instance of L; a densely packed content stream that makes everything look equally unremarkable plays the same formal role.

3.3 Ambiguity

Assumption 7. *Let x_n be observed with error ε_n, satisfying $\mathbb{E}[\varepsilon_n] = 0$ and $\mathrm{Var}(\varepsilon_n) = \sigma^2 \in [\sigma_p^2, \sigma_p^2 + \Delta]$ for some $\Delta \geq 0$ reflecting genuine uncertainty about the mechanism's own instrument. Let f_n be an external reference (e.g. a crowd consensus) satisfying $\mathbb{E}[f_n - x_n] = 0$ and $\mathrm{Var}(f_n - x_n) = \sigma_f^2$, with ε_n and $f_n - x_n$ independent. The mechanism blends $\hat{b} = (1-\eta)(x_n+\varepsilon_n)+\eta f_n$, choosing $\eta \in [0,1]$ to minimize worst-case expected squared error.*

Theorem 8 (Ambiguity-driven substitution)**.** *Under Assumption 7, the robust optimal weight is $\eta^*(\Delta) = \dfrac{\sigma_p^2 + \Delta}{\sigma_p^2 + \Delta + \sigma_f^2}$, strictly increasing in Δ, with $\eta^*(\Delta) \to 1$ as $\Delta \to \infty$.*

Proof. Since the two error sources are independent and mean zero, the worst case over the ambiguity set reduces, for each $\eta \in [0,1]$, to minimizing the convex quadratic $C(\eta) = (1-\eta)^2(\sigma_p^2 + \Delta) + \eta^2\sigma_f^2$. Its unique minimizer, by the standard first-order condition for a convex quadratic (a robust-estimation argument in the tradition of Huber and Ronchetti, 2009), is the stated formula; it lies automatically in $(0,1)$, and monotonicity in Δ follows by direct differentiation. □

Corollary 9 (Freeze under ambiguity-driven substitution)**.** *Extending $\Delta \in [0,\infty]$ so that $\eta^*(\infty) := 1$ by continuous extension of Theorem 8's formula, exact freeze under ambiguity requires both conditions jointly: $\Delta = \infty$ (unbounded self-uncertainty)* and *f_n itself frozen by one of the routes above. Unbounded self-uncertainty alone gives only $\eta^* = 1$, i.e. $\hat{b}_n \equiv f_n$: it redirects the comparison entirely onto f_n but does not by itself halt updates, since f_n may still move with x_n. Only when f_n is additionally frozen does $\hat{x}_n$ stop registering any change in the true x_n. For any finite Δ, $\eta^*(\Delta) < 1$ strictly, so $\Delta = \infty$ is a limiting idealization, not a value reached at finite uncertainty; for large finite Δ, $1-\eta^*(\Delta) = \sigma_f^2/(\sigma_p^2+\Delta+\sigma_f^2) \to 0$, so the mechanism's assessment approaches f_n arbitrarily closely without reaching it.*

3.4 Information Limit

Assumption 10. *When active, suppose the mechanism does not compare x_n against $\hat{x}_{n-1}$ via a fixed threshold, but observes a signal y_n and must decide between an unperturbed regime N and a drifted regime D, with $y_n \mid N \sim Q_N$, $y_n \mid D \sim Q_D$, equally likely a priori. A test is any rule $\delta(y) \to \{N, D\}$.*

Definition 11. *The* discriminability *of (Q_N, Q_D) is $\mathrm{Acc}^\star := \max_\delta \mathbb{P}(\delta(y) = R)$, where R is the true regime.*

Proposition 12 (Detection ceiling, total-variation form)**.** *Under Assumption 10, $\mathrm{Acc}^\star = \frac{1}{2} + \frac{1}{2}\,\mathrm{TV}(Q_N, Q_D)$, where $\mathrm{TV}(Q_N, Q_D) = \frac{1}{2}\sum_y |Q_N(y) - Q_D(y)|$.*

Proof. A test δ partitions the observation space into $S_N = \delta^{-1}(N)$, $S_D = \delta^{-1}(D)$; with a uniform prior, $\mathbb{P}(\text{correct}) = \frac{1}{2} + \frac{1}{2}[Q_N(S_N) - Q_D(S_N)]$, maximized by taking $S_N = \{y : Q_N(y) > Q_D(y)\}$, since including any point with $Q_N(y) < Q_D(y)$ strictly decreases the bracket and excluding any point with $Q_N(y) > Q_D(y)$ forfeits a positive contribution. Because $\sum_y (Q_N(y) - Q_D(y)) = 0$, the maximal bracket value equals $\mathrm{TV}(Q_N, Q_D)$ by definition. □

Proposition 13 (Detection ceiling, KL form)**.** *Under Assumption 10, $\mathrm{Acc}^\star \leq \frac{1}{2} + \frac{1}{2}\sqrt{D_{KL}(Q_N \| Q_D)/2}$.*

Proof. By Pinsker's inequality (Cover and Thomas, 2006), $\mathrm{TV}(Q_N, Q_D) \leq \sqrt{D_{KL}(Q_N \| Q_D)/2}$; substituting this bound directly into Proposition 12's identity $\mathrm{Acc}^\star = \frac{1}{2} + \frac{1}{2}\mathrm{TV}(Q_N, Q_D)$ gives $\mathrm{Acc}^\star \leq \frac{1}{2} + \frac{1}{2}\sqrt{D_{KL}(Q_N \| Q_D)/2}$ exactly, with no further approximation. □

Entropy gives a natural gloss on this ceiling. $D_{KL}(Q_N \| Q_D)$ measures how much the two regimes differ in the information they carry about which is true; a small divergence is, equivalently, a state of high uncertainty, an *entropic uncertainty*, about which regime is actually being observed. The relation runs in one direction:

$$\text{entropic uncertainty} \uparrow \Rightarrow \text{ distinguishability (TV/KL)} \downarrow \Rightarrow \text{ detection accuracy} \downarrow .$$

No new construct is introduced here: entropic uncertainty is read off the same $D_{KL}(Q_N \| Q_D)$ already defined above, not a separately defined entropy measure, and the term is used only as an interpretive label for Proposition 13's bound.

Corollary 14 (Statistical freeze under the detection ceiling)**.** *If $D_{KL}(Q_N \| Q_D) \leq 2\varepsilon^2$, then $\mathrm{Acc}^\star \leq \frac{1}{2} + \frac{\varepsilon}{2}$: correct identification of drift is capped near a coin flip, regardless of the mechanism's activity, resourcing, or self-trust.*

Remark 15. *This is a different kind of freeze than the first three: $\hat{x}_n$ may still change, but no more reliably than a coin flip once $D_{KL}(Q_N \| Q_D)$ has shrunk enough, a limit that binds on what is available to detect rather than on the detector. As $Q_D \to Q_N$, this ceiling forces $\mathrm{Acc}^\star \to \frac{1}{2}$ regardless of Sections 3.2 and 3.3, suggesting, though not establishing, that a corrector recognizing this ceiling would have independent rational grounds for the cost- or ambiguity-driven freezes of those sections.*

The Correction Freeze Theorem

Theorem 16 (Correction Freeze)**.** ***(A) Exact freeze.*** *$\hat{x}_n \equiv \hat{x}_0$ for all $n \leq N$ if any of: (i)* Threshold*, every $g_n < \tau$ (Corollary 3); (ii)* Cost*, $L_n < L^\dagger$ throughout (Corollary 6); (iii)* Ambiguity*, $\Delta_n = \infty$ throughout in the extended sense of Corollary 9, and f_n is itself frozen.* ***(B) Statistical freeze.*** *Independently, if (iv)* Information limit*, $D_{KL}(Q_N \| Q_D) \leq 2\varepsilon^2$ throughout (Corollary 14), then every active-step decision is correct with probability at most $\frac{1}{2} + \frac{\varepsilon}{2}$.*

Proof. (A) is immediate from Corollaries 3, 6, 9: each makes the update rule's correction branch fail at every step, so $\hat{x}_n$ never leaves $\hat{x}_0$. (B) is Corollary 14 applied at each active step. □

These four routes are independent in their triggering conditions: (i) needs nothing about cost or ambiguity, only small steps; (ii) needs nothing about step size, only that checking is not worth it; (iii) needs nothing about either, only that self-trust has rationally collapsed; (iv) needs none of the above, holding even for a fully active, fully resourced, fully self-trusting corrector, provided only that what it is distinguishing has become statistically close. This is what makes the theorem a genuine generalization, and what (iv) specifically closes: the one gap the first three leave open. One qualification: (iii) is independent in what triggers it, unbounded self-uncertainty alone, but it yields an exact freeze only in combination with f_n itself being frozen by one of the other three routes; absent that, substitution redirects the comparison without necessarily halting it.

4 From Freeze to Hidden Capture

4.1 The Chain from Drift to Capture

The consequence of any exact freeze is immediate: since $\hat{x}_n$ does not move while x_n can move without bound (Section 3.1), the two states diverge without limit. The consequence of interest is what happens when that divergence meets a system whose own dynamics amplify themselves, a positive feedback, or bandwagon, effect, so that the further x_n moves, the easier it becomes to move further still.

Drift → Freeze → Divergence → Feedback → Hysteresis → Hidden Capture

An increment pattern satisfying any of Section 3's four conditions (drift) forces $\hat{x}_n$ to freeze, exactly under (i)-(iii) or statistically under (iv) (freeze); under exact freeze the true state is then unconstrained by correction and can move arbitrarily far, arbitrarily far within $\mathbb{R}_{\geq 0}$ in general, or up to $1 - x_0$ within the bounded $[0, 1]$ state space used from Section 4.2 onward (Proposition 2), while under statistical freeze it is constrained only in the weak, per-step sense of Theorem 16(B) (divergence); if x_n's own dynamics reinforce themselves, made precise below (feedback), the resulting trajectory exhibits a gap between the pressure needed to trigger a shift and the smaller pressure below which it reverts (hysteresis); and a system that crosses into the shifted state while frozen completes that shift, and inherits that gap, without correction ever having had a chance to act (hidden capture). Sections 4.2 and 4.3 make this precise for the exact-freeze case; the statistical-freeze variant is discussed as a remark alongside Corollary 20.

4.2 Capture Dynamics

Assumption 17. *Let $x \in [0, 1]$ evolve as $\dot{x} = F(x; \mu) = x(1 - x)(\mu + \beta x)$, where $\mu = I\kappa - \theta$ (external inducement κ, scaled by responsiveness I, net of intrinsic resistance θ), and $\beta > 0$ is a bandwagon reinforcement parameter: the more of the state has already moved, the lower the marginal resistance to moving further.*

Proposition 18 (Trichotomy)**.** *The equilibria of F on $[0, 1]$ are $x = 0$, $x = 1$, and (for $-\beta < \mu < 0$) $x_c = -\mu/\beta \in (0, 1)$. If $\mu > 0$, $x = 1$ is the unique stable equilibrium (guaranteed capture). If $\mu < -\beta$, $x = 0$ is the unique stable equilibrium (guaranteed failure). If $-\beta < \mu < 0$, both $x = 0$ and $x = 1$ are stable and x_c is unstable (bistable regime).*

Proof. $F(x; \mu) = \mu x + (\beta - \mu)x^2 - \beta x^3$ has roots exactly at $x = 0, 1, x_c$. By the standard linearization criterion for scalar autonomous flows, stable iff $F'(x^\star) < 0$ (see e.g. Strogatz, 2015), $F'(0; \mu) = \mu$ and $F'(1; \mu) = -(\mu + \beta)$ follow directly. At x_c, writing $F(x; \mu) = x(1 - x)(\mu + \beta x)$ as a product and using $\mu + \beta x_c = 0$ by definition of x_c, the product rule collapses to $F'(x_c; \mu) = x_c(1 - x_c)\beta$, since the term differentiating the vanishing factor $(\mu + \beta x)$ drops out; substituting $x_c = -\mu/\beta$ gives $F'(x_c; \mu) = -\mu(\mu + \beta)/\beta > 0$ on the range where x_c exists, giving the three stability regions directly. Since F is a cubic with exactly these roots on $[0, 1]$, its sign cannot change between consecutive roots without a fourth root, so local stability at each equilibrium determines the flow on its entire adjacent interval. □

Proposition 19 (Hysteresis)**.** *Treating κ as slowly varying, write $\kappa_U = \theta/I$, $\kappa_L = (\theta - \beta)/I$, so $\kappa_L < \kappa_U$. Started at $x \approx 0$, increasing κ leaves the system at $x \approx 0$ until κ_U, where x jumps to $x \approx 1$. Started at $x \approx 1$, decreasing κ leaves it at $x \approx 1$ until κ_L, where it jumps back. The gap has width $\kappa_U - \kappa_L = \beta/I$.*

Proof. For $\kappa < \kappa_L$, Proposition 18 makes $x = 0$ the unique attractor. For $\kappa \in (\kappa_L, \kappa_U)$, $x = 0$ remains stable and the unstable $x_c(\kappa)$ acts as a barrier rather than an attractor, so the system stays at $x \approx 0$ throughout. At $\kappa = \kappa_U$, $F'(0; 0) = 0$ marks a saddle-node collision at which $x = 0$ loses stability as $x_c \to 0$, forcing the jump to $x \approx 1$. The reverse sweep is symmetric, colliding at κ_L where $F'(1; -\beta) = 0$. Since $\beta > 0$, $\kappa_U > \kappa_L$ strictly. □

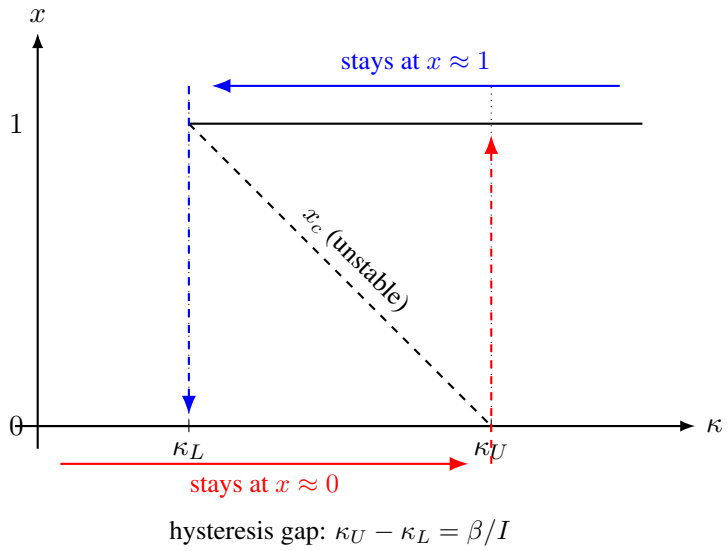

Figure 2: The hysteresis loop of Proposition 19. Sweeping κ upward from $x \approx 0$ (red) holds the system at $x \approx 0$ until κ_U, then jumps to $x \approx 1$; sweeping κ downward from $x \approx 1$ (blue) holds the system at $x \approx 1$ until κ_L, then jumps back. Any $\kappa \in (\kappa_L, \kappa_U)$ is consistent with either state, depending on history, which is exactly the gap Corollary 20 exploits: entry at κ_U and exit at κ_L are not the same threshold.

4.3 Hidden Capture

Corollary 20 (Hidden Capture)**.** *Suppose x_n evolves under Assumption 17 while any of conditions (i)-(iii) of Theorem 16(A) holds throughout, so $\hat{x}_n \equiv \hat{x}_0 \approx 0$ exactly. If κ rises past κ_U, Proposition 18 forces $x_n \to 1$, without $\hat{x}_n$ ever registering the change. If κ then falls back below κ_U but stays above κ_L, Proposition 19 shows x_n does not revert: reversal requires driving κ below κ_L specifically, a strictly harder condition than the κ_U originally crossed unnoticed. This deterministic guarantee is specific to (i)-(iii); under condition (iv) alone, the same conclusion holds only in the weaker, probabilistic sense of the remark below.*

Proof. The first claim is Proposition 18 applied to x_n's trajectory, unaffected by $\hat{x}_n$; that $\hat{x}_n$ does not register it is Theorem 16(A). The second is Proposition 19 applied from $x \approx 1$. □

Remark 21. *Under condition (iv) rather than (i)-(iii), the same conclusion holds only probabilistically: by Theorem 16(B), correction succeeds at each active step with probability at most $\frac{1}{2} + \frac{\varepsilon}{2}$, so the mechanism does no better than a biased coin at catching the approach to capture. Whether this weaker guarantee still permits x_n to cross κ_U depends on an unmodeled feedback from successful correction back into κ or x_n, so the claim is stated only as a plausible variant.*

5 Implications and Boundary Conditions

The formal results above concern an abstract true state and perceived state; nothing in Sections 2 through 4 names a real system, and confirming that any real case satisfies the stated hypotheses would require measuring quantities, an increment size, a verification cost, a divergence, a bandwagon parameter, that this paper has only treated as symbols. Read as interpretation rather than measurement, the four routes give a precise vocabulary for otherwise loose worries about digital life. A habit such as smartphone checking can rise through increments too small to notice (Section 3.1), while a bandwagon-reinforced norm makes it harder to undo than to form (Section 4.2's hysteresis gap). Synthetic content competing with genuine content

in the same feeds can defeat verification through no one's carelessness but through the information limit of Section 3.4, once the two become statistically close enough that no vigilance restores the gap. And wherever the volume of claims to evaluate outpaces a roughly fixed attention budget, Section 3.2's cost mechanism predicts not a gradual thinning of scrutiny but a sharp collapse to none at all, the pattern popularly called information overload. A structurally similar account appears independently in mindsponge theory (Vuong, 2023), which frames such filtering as an energy-conserving strategy under survival pressure, a framing with a concrete grounding in cell biology: usable energy sets a hard ceiling on how much internal correction a system, biological or social, can sustain (Lane, 2015).

A further, and more recent, instance is overuse of AI tools themselves. The share of US adults who have used ChatGPT roughly doubled between 2023 and 2025 and had reached about half of all adults by mid-2026 (Sidoti and McClain, 2025; Gottfried et al., 2026), and a 2026 study of working adults in Shanghai found that AI dependency, compulsive use, fear of missing out, and anxiety each independently predicted overreliance, an uncritical acceptance of AI output that includes reduced verification of what the tool produces, with cognitive overload as a partial mediator (Zhang et al., 2026). Read through Section 3.3, this is ambiguity-driven substitution at societal scale: the more routinely a person offloads a judgment to an AI system, the less practiced, and so the less trusted, their own unaided judgment becomes, which is exactly the growing Δ that makes further substitution the rational response rather than a lapse. Because that erosion of self-trust is itself produced by the substitution it justifies, the pattern also matches Section 4.2's reinforcement structure, giving one reason to expect that reducing an already-established reliance on AI takes more deliberate effort than avoiding it would have.

Three caveats apply throughout. The composition is mathematically valid only through its stated hypotheses; it does not establish that any real system's τ, $L^{\dagger}$, Δ, or D_{KL} are measurable, independent quantities rather than constructs chosen to make the composition go through. The correction mechanism is assumed to compare only against the immediately preceding true value x_{n-1}, and I, θ, β are treated as exogenous constants; richer memory models and endogenous parameters are natural extensions left open. Most importantly, nothing here is fit to data or intended as a predictive claim about any actual person, market, or population; the results are offered for orientation, a precise account of what follows from a stated set of assumptions. Every result above is proved using standard tools, a telescoping sum, a convex first-order condition, a linearized stability check, a named divergence inequality, so that no step depends on a source the reader could not independently check.

6 Conclusion

The correction freeze problem, stated as a question, is whether a system's belief about itself can become detached from what is actually true without any single step ever registering as a violation. This paper answers it in four independent ways, a threshold effect, a cost effect, an ambiguity effect, and an information limit, each sufficient on its own to freeze correction, exactly, or, in the fourth case, statistically. The consequence is that the true state can continue moving unseen, without bound in $\mathbb{R}_{\geq 0}$, or up to the domain boundary in the bounded $[0, 1]$ setting used from Section 4.2 onward (Proposition 2). And where that divergence meets a system whose own dynamics reinforce themselves, the hysteresis gap between entering and leaving a shifted state turns an unnoticed drift into a hidden, and disproportionately difficult to reverse, capture. That hierarchy, problem, mechanisms, consequence, dynamics, is the paper's central claim; everything else is elaboration.

Acknowledgment

I am grateful to Dr. Nguyen Minh Hoang of Phenikaa University for his thoughtful discussions during the preparation of this paper.

References

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley.

Gottfried, J., Bishop, W., Anderson, M., Faverio, M., Park, E., & McClain, C. (2026). *Americans and AI 2026: Chatbots, smart devices and views on impact*. Pew Research Center.

Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics* (2nd ed.). Wiley.

Lane, N. (2015). *The vital question: Energy, evolution, and the origins of complex life*. W. W. Norton & Company.

Sidoti, O., & McClain, C. (2025). *34% of U.S. adults have used ChatGPT, about double the share in 2023*. Pew Research Center.

Strogatz, S. H. (2015). *Nonlinear dynamics and chaos* (2nd ed.). Westview Press.

Vuong, Q.-H. (2023). *Mindsponge theory*. Sciendo.

Zhang, J., Yao, Z., Kang, D., Rasiah, R., & Gao, N. (2026). A SEM-ANN analysis to examine impact of AI overreliance through a social cognitive lens: The roles of dependency, FOMO, addiction, and anxiety. *BMC Psychology*, *14*, Article 547.

Appendix: Table of Mathematical Symbols

Symbol	Meaning	Where used
n	discrete time index	Section 2
x_n	true state at step n	Section 2
$\hat{x}_n$	perceived state at step n	Section 2
g_n	increment of the true state, $g_n = x_n - x_{n-1}$	Section 2
τ	perceptibility threshold in the update rule	Section 2
D	target cumulative distortion	Section 3.1
N	number of steps needed to reach D	Section 3.1
k	verification effort	Section 3.2
c	marginal cost of one unit of verification effort	Section 3.2
$\Lambda(k)$	expected residual loss remaining after k units of checking	Section 3.2
Λ_0, ρ	constants of the exponential loss function	Section 3.2
L	visible-range variable	Section 3.2
$g(L)$	perceived-stakes multiplier, strictly increasing in L	Section 3.2
$C(k)$	total verification cost	Section 3.2
$k^*(L)$	cost-minimizing verification level	Section 3.2
$L^\dagger$	threshold visible range below which $k^*(L) = 0$	Section 3.2
ε_n	observation error on x_n	Section 3.3
σ^2	variance of the observation error	Section 3.3
σ_p^2, Δ	bounds of the ambiguity set, $\sigma^2 \in [\sigma_p^2, \sigma_p^2 + \Delta]$	Section 3.3
f_n	external reference signal	Section 3.3
σ_f^2	variance of the error in f_n	Section 3.3
$\hat{b}$	blended estimate of x_n	Section 3.3
η	blending weight on the reference signal	Section 3.3
$\eta^*(\Delta)$	robust optimal blending weight	Section 3.3
y_n	observed signal used to distinguish regimes	Section 3.4
N, D (regimes)	unperturbed regime and drifted regime	Section 3.4
Q_N, Q_D	signal distributions under regime N and regime D	Section 3.4
δ	a test, or decision rule, mapping a signal to a regime	Section 3.4
R	the true regime	Section 3.4
$\mathrm{Acc}^\star$	best accuracy achievable by any test	Section 3.4
$\mathrm{TV}(Q_N, Q_D)$	total variation distance between Q_N and Q_D	Section 3.4
S_N, S_D	partition of the observation space induced by a test	Section 3.4
$D_{KL}(Q_N \Vert Q_D)$	Kullback-Leibler divergence between regime distributions	Section 3.4
ε	detection-ceiling accuracy margin	Section 3.4
x	continuous-time population capture share	Section 4.2
μ	net driving pressure, $\mu = I\kappa - \theta$	Section 4.2
I	responsiveness to the external inducement	Section 4.2
κ	external inducement, or capacity, applied to the population	Section 4.2
θ	intrinsic resistance of the population	Section 4.2
β	reinforcement, or bandwagon, parameter	Section 4.2
$F(x; \mu)$	right-hand side of the capture dynamics, $\dot{x} = F(x; \mu)$	Section 4.2
x_c	interior, unstable equilibrium of F	Section 4.2
κ_U, κ_L	upper and lower capacity thresholds for capture and reversal	Section 4.2, Section 4.3