Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Structural Crystallization: A Unified Computational Framework for Memory Formation, Persistence, and Modification

Po-Ting Lin

botimlin@gmail.com

independent researcher https://orcid.org/0009-0002-4825-6905

Research Article

Keywords: memory consolidation, reconsolidation, fear conditioning, computational modeling

Posted Date: April 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9387132/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Structural Crystallization: A Unified Computational Framework for Memory Formation, Persistence, and Modification

Po-Ting Lin

Independent Researcher

Author Note

Po-Ting Lin https://orcid.org/0009-0002-4825-6905

Correspondence concerning this article should be addressed to Po-Ting Lin. E-mail: botimlin@gmail.com

Author Contributions (CRediT): Po-Ting Lin: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing—original draft, Writing—review & editing.

Abstract

Prevailing models of memory treat strength as a unitary quantity that increases with learning and decreases with forgetting or extinction. We argue that this conflation obscures a fundamental distinction: memory has separable dimensions of structural accumulation and representational fidelity, which are differentially modified by retrieval. Structural accumulation captures how much of a memory trace has been consolidated; representational fidelity captures how faithfully that trace preserves its original encoding. The two can change independently—a memory can retain its full structural extent while becoming progressively distorted, or dissolve while remaining perfectly faithful. This distinction, combined with a second distinction between structurally protective retrieval (externally guided, high-constraint) and structurally risky retrieval (internally generated, low-constraint), resolves several phenomena that have resisted unified explanation: why extinction is temporary but retrieval-extinction can produce lasting change, why stress-enhanced fear memories persist for orders of magnitude longer than ordinary memories, and why fear extinction is fragile while appetitive extinction is durable. We formalize these distinctions in a system of four coupled differential equations (the Structural Crystallization framework), calibrate it to two datasets, and show that it correctly predicts—without parameter adjustment—outcomes across five independent benchmarks where competing models (Rescorla-Wagner; latent cause) fail. Extensions to human declarative memory capture the testing effect crossover, while an explicit failure on the spacing effect reveals interpretable boundary conditions. The results demonstrate that separating accumulation from fidelity is not merely a modeling convenience but a theoretical necessity for any account of both the persistence and the modifiability of memory.

Keywords: memory consolidation, reconsolidation, fear conditioning, computational modeling

Structural Crystallization: A Unified Computational Framework for Memory Formation, Persistence, and Modification

Introduction

The Conflation Problem

A memory can be strong yet wrong. A trauma survivor may vividly and confidently recall details of an event that, when checked against records, turn out to be substantially distorted (Talarico & Rubin, 2003). The memory feels absolutely real; much of its content is fabricated or borrowed from later retellings. Strength and accuracy have come apart.

This dissociation is not limited to exceptional events. A patient who has undergone successful exposure therapy for a specific phobia may show no fear for months, only to relapse when tested in a new context or after sufficient time has passed (Bouton, 2004; Craske et al., 2008). The therapeutic memory was genuinely formed—the patient was not pretending—but it dissolved while the original fear memory persisted. Two memories formed under different conditions decayed at different rates, and the one formed under lower emotional arousal lost the competition.

Perhaps most striking is the case in which the same act of remembering produces opposite outcomes depending on timing. In the retrieval-extinction paradigm (Monfils et al., 2009; Schiller et al., 2010), a brief reminder of a conditioned fear stimulus, followed by extinction training within a narrow temporal window, produces lasting fear reduction that resists spontaneous recovery, reinstatement, and renewal. The same reminder followed by the same extinction training outside this window produces only standard, temporary extinction. The organism performed the same behavior—it recalled the fear memory, then underwent extinction—but the structural consequences for the memory trace were qualitatively different.

These three observations—a memory that is strong but inaccurate, a therapeutic memory that forms but dissolves, and an identical procedure that permanently modifies or merely suppresses depending on timing—share a common structure. Each reveals a

dissociation that any framework treating memory as a single quantity cannot capture. If memory strength is one number, then a strong memory is implicitly an accurate memory, a dissolving memory is implicitly a weakening memory, and retrieval either strengthens or weakens but cannot selectively alter content while preserving structure. The empirical record says otherwise.

Yet the dominant computational frameworks in learning and memory operate precisely under this single-strength assumption. In the Rescorla-Wagner model (Rescorla & Wagner, 1972) and its descendants (Pearce & Hall, 1980), a conditioned stimulus carries one associative strength that increases with reinforcement and decreases with extinction. In latent cause models (Gershman et al., 2017), each inferred cause carries a single set of associative weights. In both frameworks, a memory that is "strong" is implicitly also "accurate"—there is no variable that tracks whether the stored representation still resembles what was originally encoded, and no variable that tracks whether the structure has been deformed by the act of retrieval.

We argue that this conflation is not a minor simplification but a representational impossibility. A single scalar cannot simultaneously encode two logically independent properties: how much of a trace exists and how faithfully that trace preserves its original content. These are different kinds of information. The first is a quantity—it answers "how much?" The second is a relation between the current state and the original state—it answers "how changed?" Collapsing them into one number forces the framework to treat every reduction in behavioral expression as the same kind of event, whether it arose from dissolution of structure, distortion of content, or competition from another trace. The empirical record requires distinguishing among these, and no single-scalar representation can do so. The problem is not that existing models have the wrong parameters; it is that they have the wrong number of dimensions.

The core claim of this article is:

> Memory is not scalar but composite. It consists of at least two independent

> dimensions—structural accumulation and representational fidelity—which are differentially modified by retrieval. Any framework that collapses these dimensions into a single quantity is subject to a representational constraint that prevents it from simultaneously accounting for both the persistence and the modifiability of memory.

Structural accumulation captures how much of a memory trace has been consolidated—the existence of structure. Representational fidelity captures how faithfully that structure preserves its original encoding—the integrity of what exists. These two dimensions can change independently. A memory can retain its full structural extent while becoming progressively distorted (high accumulation, low fidelity), or it can dissolve while remaining perfectly faithful to its original form (declining accumulation, intact fidelity). The behaviorally observable output—what the organism expresses—depends on both: effective memory strength is the product of how much structure exists and how faithfully that structure is preserved.

Existing Frameworks and Their Structural Constraints

The claim that accumulation and fidelity must be separated is strong: it asserts not merely that the separation is useful, but that it is *necessary*—that no framework lacking this distinction can provide a unified account of the phenomena listed above. To substantiate this claim, we examine the structural constraints that existing frameworks face.

Single-Strength Models

The Rescorla-Wagner model (Rescorla & Wagner, 1972) provides the foundational account of Pavlovian conditioning: associative strength V increases when a reinforcer exceeds expectation and decreases when expectation exceeds outcome. This single-variable framework elegantly captures acquisition, extinction, blocking, and overshadowing within a common error-correction rule.

However, a single associative strength cannot distinguish between a memory that has been *weakened* (less structure exists) and a memory that has been *distorted* (the same amount of structure exists, but it no longer faithfully represents the original encoding). In the Rescorla-Wagner framework, both situations manifest identically as a reduction in V. This means the model has no mechanism for reconsolidation effects: a retrieval event before extinction cannot differentially affect V, because there is no variable that tracks structural fragility and no variable that tracks deformation. The model predicts that retrieval-extinction and standard extinction produce identical outcomes—contrary to the robust finding that they do not (Monfils et al., 2009; Schiller et al., 2010).

This limitation is not parametric but structural. No choice of learning rates, decay constants, or additional associative strengths within the single-strength architecture can produce a mechanism by which retrieval opens a temporal window during which subsequent learning modifies the original trace rather than creating a competing one. The architecture lacks the representational capacity to distinguish "the trace is weaker" from "the trace is altered."

Latent Cause Models

The latent cause model of Gershman et al. (2017) introduces a richer representational structure by positing that the organism infers which latent cause generated current observations, using a distance-dependent Chinese Restaurant Process for cause assignment and Rescorla-Wagner learning within each cause. This architecture naturally accommodates reconsolidation: when a retrieval event reactivates the original cause, subsequent extinction modifies the cause directly rather than creating a competing one.

With its default parameters (not fitted to any dataset), this model successfully predicts retrieval-extinction blocking (Schiller et al., 2010), an interval-dependent gradient (Monfils et al., 2009), and within-subject dissociations between reminded and non-reminded stimuli. These successes reflect the model's core insight: cause-reactivation determines whether extinction updates or replaces.

However, when the model is constrained to fit empirical time courses—specifically, the multi-day spontaneous recovery curve reported by Quirk (2002)—its out-of-sample predictions collapse. As we demonstrate in the Results, the best-fitting parameters (RMSE = 0.050, the lowest of all models tested) produce a Schiller blocking effect that disappears (retrieval-extinction recovery: 0.35 vs. standard: 0.42) and a Monfils interval gradient that *inverts* (short intervals produce *higher* recovery than long intervals). The time-scale parameter required to reproduce the days-to-weeks dynamics of spontaneous recovery disrupts the trial-to-trial cause-switching dynamics responsible for reconsolidation predictions.

This failure is instructive. It reveals that within the latent cause architecture, the temporal dynamics governing spontaneous recovery (a slow process operating over days) and the temporal dynamics governing reconsolidation (a fast process operating over hours) are coupled through the same parameters. The framework cannot simultaneously satisfy both constraints because it has one set of temporal dynamics, not two. As we will argue, two sets of temporal dynamics—one for accumulation, one for fidelity—are precisely what the accumulation–fidelity distinction provides.

Storage Strength and Retrieval Strength

The distinction we propose has a conceptual predecessor. Bjork and Bjork (1992) proposed a new theory of disuse in which each memory item carries two strengths: storage strength (how well learned the item is, which only increases) and retrieval strength (how accessible the item is, which fluctuates with recency and context). This theory provided a principled account of desirable difficulties—conditions under which making learning harder in the short term enhances retention in the long term—and has been influential in educational psychology.

The accumulation–fidelity distinction is related to but not equivalent to the storage–retrieval distinction. Crystallization progress D is closer to storage strength in that it captures how much structure has been consolidated. However, D can decrease through

dissolution, unlike Bjork and Bjork's storage strength, which monotonically increases. Fidelity loss F is not equivalent to retrieval strength: F captures structural *distortion*—actual degradation of the trace's content—not accessibility or retrieval difficulty. Critically, F does not recover spontaneously. Once a memory's structure has been deformed, it remains deformed unless corrected by externally guided re-exposure to the original encoding conditions. This permanence gives the framework its account of lasting change through retrieval-extinction: the fear trace is not suppressed by a competing memory but structurally altered.

The most important difference is formal. The new theory of disuse was proposed as a verbal-conceptual framework and has never been instantiated as a dynamic system with coupled differential equations. Its predictions remain qualitative and difficult to test against alternative accounts. The present work can be viewed, in part, as a formalization of the intuition behind the storage–retrieval separation, extended with explicit mechanisms for labilization, deformation, and arousal-dependent persistence, and demonstrated to be computationally sufficient for phenomena that the verbal theory cannot adjudicate.

Contextual Retrieval

Bouton (1993, 2004) proposed that extinction does not erase the original conditioning but creates a new, context-dependent inhibitory memory. When context changes—through passage of time (spontaneous recovery), a shift in physical environment (renewal), or an unsignaled presentation of the US (reinstatement)—the inhibitory memory loses its retrieval advantage and the original conditioning is expressed. This account is well supported by evidence that extinction is context-specific while conditioning is not.

The Structural Crystallization framework is compatible with this account at the level of the safety pathway: D_{safety} is indeed a competing trace whose expression can be context-dependent and time-limited. The framework adds a second mechanism—fidelity loss (F)—that operates directly on the fear trace and is neither context-dependent nor time-limited. This second mechanism is what distinguishes standard extinction (which

relies entirely on the competing safety trace and is therefore vulnerable to the recovery phenomena Bouton identified) from retrieval-extinction (which additionally deforms the fear trace itself, producing durable change that resists context shifts). Bouton's account explains why standard extinction fails; the accumulation–fidelity distinction explains why retrieval-extinction can succeed.

Memory Reconsolidation

The discovery that consolidated memories, when reactivated, return to a labile state requiring protein synthesis for restabilization (Nader et al., 2000) revitalized interest in the modifiability of established memories. Subsequent work established that the reconsolidation window has a characteristic time course, that different types of learning can interfere with restabilization, and that the boundary conditions for reconsolidation induction involve prediction error, memory age, and trace strength (Lee et al., 2017).

The Structural Crystallization framework formalizes the reconsolidation concept through the labilization variable R: a transient state induced by unreferenced access that decays exponentially with time constant τ_R. The framework adds a mechanistic distinction that the reconsolidation literature has not computationally formalized: the separation between *labilization* ($R > 0$: the structure is fragile) and *modification* (F increases: the structure has been deformed). In the framework, a labile memory that receives no further signal ($S = 0$) restabilizes without modification—R decays, F remains unchanged. Modification requires both lability *and* signal: the structure must be fragile, and a force must pass through it. This is consistent with evidence that reconsolidation blockade requires both reactivation and a destabilizing agent; reactivation alone is not sufficient (Lee et al., 2017).

The Common Gap

Across these frameworks, a pattern emerges. Each captures some aspect of memory dynamics but lacks the representational capacity to simultaneously account for persistence, modification, and the interaction between them:

- Rescorla-Wagner cannot distinguish weakening from distortion.
- Latent cause models can distinguish modification from new learning, but their temporal dynamics for reconsolidation and spontaneous recovery are coupled and cannot be independently satisfied.
- The new theory of disuse separates storage from retrieval conceptually but has not been formalized and does not capture structural distortion.
- Contextual retrieval explains context-dependent relapse but not lasting modification of the original trace.
- Reconsolidation theory describes labilization but does not computationally distinguish it from modification.

The pattern across these frameworks is not a coincidence of individual modeling choices. It reflects a deeper constraint: any theory that represents memory state as a single quantity—whether that quantity is called associative strength, posterior probability, retrieval strength, or activation level—is subject to a representational bottleneck. A single variable must simultaneously encode the amount of structure that exists, the fidelity of that structure to its original form, and the trace's current vulnerability to modification. These are logically independent properties, and forcing them into one dimension creates trade-offs that manifest as the specific failures catalogued above. The missing element is not a better single variable but a second dimension: a formal separation between structural existence and representational integrity, combined with a mechanism that specifies when retrieval is protective and when it is destructive.

Two Retrieval Modes, Two Consequences

The accumulation–fidelity distinction gains explanatory power when combined with a second distinction: between two modes of memory access that have opposite effects on the trace.

When a memory is accessed through rich external cues that constrain activation across many features simultaneously—what we term *referenced access*—the retrieval is

structurally protective. Multiple constraints cross-check one another, preventing any feature from being driven beyond its stable range. Referenced access strengthens the trace (advances accumulation) and corrects prior distortions (repairs fidelity).

When a memory is reconstructed internally from sparse cues—*unreferenced access*—the retrieval is structurally risky. Few constraints mean the activation is poorly controlled. Features that are not constrained may be driven beyond tolerance and deform (*overshoot*), or may fail to be activated at all (*incomplete readout*). Unreferenced access can advance accumulation (the act of reconstruction is itself a learning event), but it also renders the trace temporarily fragile, and any signal passing through the fragile trace at low constraint density degrades fidelity.

The combination of these two distinctions—accumulation versus fidelity, referenced versus unreferenced access—generates a simple but powerful logic:

- *Standard extinction* builds a new safety trace (increases safety accumulation) but does not touch the fear trace's fidelity. The safety trace dissolves over time. Fear returns.

- *Retrieval-extinction* first renders the fear trace fragile through unreferenced access, then passes signals through it during extinction. Fidelity degrades permanently. Fear does not return—not because a safety trace suppresses it, but because the fear trace itself has been structurally deformed.

- *Stress-enhanced fear learning* locks a high arousal parameter at encoding that exponentially suppresses dissolution. The memory is not "deeper"—initial accumulation is similar across stress levels—but "harder": the arousal lock makes it resistant to the passage of time.

- *Appetitive extinction* is durable because both the appetitive trace and the safety trace form at the same low arousal. Neither has a dissolution advantage. They age together and the competitive balance is maintained.

None of these explanations require the organism to learn anything new about the relationship between stimuli. They follow from the structural properties of the trace itself.

The Present Work

We formalize the accumulation–fidelity distinction in the *Structural Crystallization* framework, a system of four coupled ordinary differential equations. Each memory is modeled as a multi-branch structure that crystallizes incrementally. A single permanent property—emotional arousal at encoding—determines both the speed of crystallization and the structure's resistance to subsequent dissolution.

The theoretical propositions instantiated by the framework are:

1. Memory has separable dimensions of structural accumulation (D) and representational fidelity (F). Effective memory strength is $D(1 - F)$.

2. Referenced access (externally guided, high-constraint retrieval) advances D and repairs F without inducing labilization.

3. Unreferenced access (internally generated, low-constraint retrieval) advances D but induces labilization (R), which drives F increase when constraint density is low.

4. Emotional arousal at encoding (M_i) is permanently locked at formation and exponentially suppresses subsequent dissolution of D.

5. Behavioral expression is determined by competition between the effective strengths of the relevant pathways.

The framework is calibrated to two datasets—spontaneous recovery (Quirk, 2002) and the retrieval-extinction interval gradient (Monfils et al., 2009)—and then tested, without parameter adjustment, against five independent empirical benchmarks: retrieval-extinction blocking (Schiller et al., 2010), within-subject dissociation between reminded and non-reminded stimuli, stress-enhanced fear learning persistence (Rau & Fanselow, 2009), appetitive extinction durability (Woods & Bouton, 2008), and the testing

effect crossover in human declarative memory (Roediger & Karpicke, 2006). It is compared against Rescorla-Wagner and the latent cause model under identical calibration-then-predict conditions. Extensions to retrieval-induced forgetting (Anderson et al., 1994) and flashbulb memory (Talarico & Rubin, 2003) provide further theoretical leverage, while an explicit failure on the spacing effect delineates the framework's boundary conditions.

The contribution of this article is not the model per se but the theoretical claim it instantiates: that accumulation and fidelity are separable dimensions of memory, that retrieval differentially modifies them, and that this distinction is required for any unified account of memory persistence and modification. The model serves as a formal proof of concept—a demonstration that the claim is not merely plausible but computationally sufficient.

Model Description

Overview and Physical Metaphor

The Structural Crystallization framework models each memory as a multi-branch trunk. Each branch corresponds to a constituent feature of the encoded experience—visual, auditory, emotional, spatial, temporal, or semantic. Learning proceeds by crystallizing branches one at a time onto the trunk; once a branch passes a completion threshold, it advances the trunk's overall crystallization. The trunk's behavioral expression depends on both how much structure has formed and how faithfully that structure preserves the original encoding.

The crystallization metaphor is chosen deliberately. A crystal grows incrementally, branch by branch, under constraints imposed by its existing structure. The resulting object has two independent properties that can vary separately: its *extent* (how much crystal has formed) and its *fidelity* (how closely the crystal matches its ideal lattice). A crystal can be large but full of defects, or small but perfectly ordered. Similarly, a memory can be extensively consolidated yet substantially distorted, or minimally consolidated yet perfectly

faithful. The metaphor generates the D/F separation naturally, rather than imposing it as an ad hoc modeling choice.

Four state variables describe each memory pathway i:

- $D_i \in [0, 1]$: *Crystallization progress*—the proportion of the trunk that has been structurally consolidated (global, persistent).
- $C_i \in [0, 1]$: *Branch progress*—the current branch under construction (local, transient).
- $R_i \in [0, \infty)$: *Labilization state*—structural fragility following internally generated retrieval (transient).
- $F_i \in [0, 1]$: *Fidelity loss*—cumulative deformation of the crystallized structure relative to its original configuration (global, persistent).

Each trunk also carries one permanent property: M_i, the emotional arousal level at the time of formation. M_i is set once and locked; it governs both the speed of initial crystallization and the trunk's resistance to subsequent dissolution.

The only behaviorally observable quantity is the *effective crystallization strength*, $D_i(1 - F_i)$. D_i captures how much structure exists; F_i captures how much of that structure has been distorted. C_i and R_i are internal process variables with no direct behavioral signature.

Two Access Modes

A central distinction in the framework is between two modes of memory access, which produce qualitatively different consequences for the memory trace.

Referenced Access (S_i^{ext})

When external stimuli provide cues across multiple branches simultaneously, the memory is accessed under high *branch coverage*. Multiple constraints converge on the trunk from different directions, cross-checking one another and preventing any single branch from being driven beyond its structural tolerance. This mode strengthens the

existing structure (advances C), corrects prior deformation (reduces F), and critically, does not induce labilization (R remains unchanged).

Unreferenced Access (S_i^{int})

When retrieval is internally generated from few cues, the memory is reconstructed under low branch coverage. Only a small subset of branches serves as entry points; the remaining branches must be reconstructed without external constraint. This poorly controlled activation produces two distinct failure modes.

Overshoot occurs when the retrieval signal exceeds what uncovered branches can structurally tolerate. With few constraints to calibrate the signal's magnitude, uncovered branches are driven beyond their stable range and their structure shifts—the crystallized configuration is physically displaced from its original form. This is an active process: the signal passes through the structure and pushes it out of alignment.

Incomplete readout occurs when the retrieval signal is insufficient to activate all branches. Uncovered branches simply fail to be reached, and the reconstruction that emerges is a partial, impoverished version of the original memory. This is not merely "failing to remember"—the incompleteness directly affects the memory's effective expression, because the behavioral output depends on the integrated activity across all branches, not just the ones that happen to be activated.

Both failure modes share a single root cause: low η. When coverage is low, uncovered branches are either actively deformed by an uncontrolled signal (overshoot) or passively absent from the reconstruction (incomplete readout). The net effect in both cases is a reduction in the memory's effective behavioral expression. In the fidelity loss equation (Equation 3), the term $(1 - \eta)$ captures both mechanisms simultaneously: it quantifies the proportion of branches exposed to uncontrolled activation or omission.

This mode advances C (the act of reconstruction still constitutes a learning event), but also induces labilization (R increases), which in turn drives fidelity loss (F increases) at low coverage.

Branch Coverage (η)

Branch coverage $\eta_i \in [0, 1]$ represents the proportion of branches constrained during a given access event. It is not a free parameter but a scenario setting determined directly by the access mode:

- S^{ext}: $\eta \approx 0.9$ (high coverage; external stimuli activate most branches).
- S^{int}: $\eta \approx 0.1$–0.2 (low coverage; few internal cues).
- Retrieval trial ($S^{\text{ext}} + S^{\text{int}}$): $\eta \approx 0.2$ (a single conditioned stimulus provides one branch as entry point).

The asymmetry between access modes generates a triple protection–triple exposure contrast: referenced access does not trigger R, operates at high η, and corrects F; unreferenced access triggers R, operates at low η, and does not correct F.

Governing Equations

The preceding distinctions—between accumulation and fidelity, between referenced and unreferenced access—are formalized in four coupled ordinary differential equations. Each equation governs one state variable; together, they constitute the complete dynamics of the framework.

Equation 1: Branch Progress (C)

Branch progress captures the local construction process: how far the current branch has advanced toward completion. Any access event—whether referenced or unreferenced—pushes C toward ceiling, but the rate depends on arousal and the available bandwidth.

$$\frac{dC_i}{dt} = \alpha_{C0} \cdot M \cdot \left(S_i^{\text{ext}}(t) + S_i^{\text{int}}(t)\right) \cdot (1 - C_i) - \delta_C \cdot C_i - \gamma \cdot h_i(t) \cdot C_i \tag{1}$$

where $h_i(t) = \sum_{j \neq i}(S_j^{\text{ext}}(t) + S_j^{\text{int}}(t))$ captures inter-pathway interference, and M (without subscript) is the current emotional arousal level. The first term drives crystallization,

bounded by the bandwidth limit $(1 - C_i)$: as C approaches ceiling, each additional stimulus event produces diminishing returns. The second term represents natural cooling of the branch under construction. The third term captures interference from concurrent activity on other pathways.

Equation 2: Labilization (R)

Labilization captures the transient structural fragility that follows unreferenced access. It is the formal counterpart of the reconsolidation window.

$$\frac{dR_i}{dt} = \rho \cdot S_i^{\text{int}}(t) - \frac{R_i}{\tau_R} \tag{2}$$

Only unreferenced access drives labilization. Referenced access, with its multi-branch convergence, provides sufficient constraint to prevent structural destabilization. R decays exponentially with time constant τ_R, which corresponds to the reconsolidation window.

Equation 3: Fidelity Loss (F)

Fidelity loss captures the cumulative, permanent deformation of the crystallized structure. It is the variable that distinguishes the present framework from all single-strength models.

$$\frac{dF_i}{dt} = \phi \cdot R_i \cdot \left(S_i^{\text{ext}}(t) + S_i^{\text{int}}(t)\right) \cdot (1 - \eta_i) \cdot (1 - F_i) - \psi \cdot S_i^{\text{ext}}(t) \cdot F_i \tag{3}$$

The deformation term (first term) requires four simultaneous conditions: the structure must be labile ($R > 0$), a signal must be passing through ($S > 0$), branches must be uncovered ($\eta < 1$), and deformation capacity must remain ($F < 1$). Critically, R represents structural fragility and S represents force—a labile structure that receives no signal ($R > 0$, $S = 0$) does not deform. The factor $(1 - \eta)$ captures both overshoot and incomplete readout: uncovered branches are either driven beyond tolerance or reconstructed with inaccurate fill-in, and both contribute to fidelity loss.

The repair term (second term) operates only through referenced access on the pathway's own external signal S_i^{ext}. External stimulation of a different pathway cannot correct deformation on pathway i—a safety signal is not the correct template for a fear memory's crystalline structure. The repair term has no natural recovery component: once deformation occurs, it is permanent unless the pathway receives its own referenced input.

Equation 4: Crystallization Progress (D)

Crystallization progress captures the global, persistent consolidation of the memory trunk. It advances when branch construction crosses a threshold, and dissolves at a rate that depends on the trunk's emotional arousal at formation.

$$\frac{dD_i}{dt} = \kappa \cdot \sigma\left(\frac{C_i - C_{\text{thresh}}}{w}\right) \cdot (1 - D_i) - \beta_0 \cdot e^{-D_i \cdot M_i / D_c} \cdot D_i \tag{4}$$

where $\sigma(x) = 1/(1 + e^{-x})$. The advance term is gated: C must exceed a threshold C_{thresh} before the trunk progresses, implementing the branch-by-branch sequential crystallization. The dissolution term is the framework's account of forgetting. The exponential factor $e^{-D_i \cdot M_i / D_c}$ makes dissolution self-limiting: as D increases and as the formation arousal M_i increases, the effective dissolution rate drops toward zero. This produces the key asymmetry that drives spontaneous recovery: a fear memory formed at high M_i has a half-life orders of magnitude longer than a safety memory formed at low M_i.

Labilization R does not appear in Equation 4. The effects of unreferenced access on memory strength operate exclusively through fidelity loss F, not through direct erosion of crystallized structure.

Equation 5: Resource Constraint

$$\sum_i D_i(t) \leq B \tag{5}$$

A global resource budget B limits total crystallization across all concurrent pathways. This constraint produces competition: when one pathway crystallizes rapidly

(e.g., emotional branches during trauma), it consumes resources that would otherwise support crystallization of other branches (e.g., spatial, temporal), accounting for the fragmented nature of traumatic memories.

Behavioral Readout: Expressed Fear

$$\text{expressed fear} = \sigma\left(\frac{D_{\text{fear}} \cdot (1 - F_{\text{fear}}) - D_{\text{safety}}}{\tau}\right) \tag{6}$$

Fear expression is determined by the competition between the effective strength of the fear pathway and the safety pathway. The sigmoid with temperature τ maps this competition to a $[0, 1]$ behavioral output. Within any single experiment, all test groups experience the same conditioned stimulus under the same conditions, so η at test is constant across groups and absorbed into τ. Between-group differences arise entirely from the dynamics of D and F.

Parameters

The framework contains three classes of parameters (Table 1).

Three parameters (C_{thresh}, w, ρ) were fixed *a priori* based on theoretical considerations. Nine model parameters and one scenario parameter (M_{cond}, the emotional arousal level specific to the Quirk conditioning protocol) were jointly optimized, for a total of 10 free parameters in the calibration. Three scenario parameters (M_{ext}, B, γ) were set to default values and held constant across all simulations. All out-of-sample predictions used the calibrated parameter set without adjustment. A full sensitivity and identifiability analysis is reported in Appendix B.

Method

Calibration Strategy

The framework was jointly calibrated to two datasets: the spontaneous recovery time course from (Quirk, 2002) and the retrieval-extinction interval gradient from (Monfils et al., 2009). These two datasets were chosen because they isolate complementary

mechanisms within the framework: spontaneous recovery constrains the crystallization–dissolution dynamics (Equations 1–4), while the interval gradient constrains the deformation dynamics (Equations 2–3).

Calibration followed a pattern-based strategy. For the Quirk spontaneous recovery curve (6 time points spanning days 1–14 post-extinction), the objective was soft mean squared error with additional constraints enforcing monotonic increase and plausible range. For the Monfils interval gradient (four retrieval-extinction intervals: 10 min, 1 hr, 6 hr, 24 hr), the objective imposed hard constraints: the gradient must be monotonically increasing (10 min $<$ 1 hr $<$ 6 hr $<$ 24 hr), the spread across conditions must exceed 0.15, and the shortest interval must produce spontaneous recovery below 0.5. Each violated monotonicity or spread constraint added a penalty of 10; the short-interval ceiling constraint added a penalty of 5.

Target values for both datasets were extracted visually from published figures rather than reported numerically in text, introducing unavoidable reading imprecision. Fitting to precise numerical targets under these conditions would imply a degree of measurement accuracy that the source data do not support. The pattern-based strategy—fitting qualitative shape and ordering rather than exact values—reflects this constraint. It is also consistent with the principle that different laboratories using different procedures will produce different absolute values but the same qualitative patterns (cf. Bouton, 2004).

The combined loss function was minimized using differential evolution (Storn & Price, 1997) with a population size of 20 over 100 iterations, parallelized across 16 workers. Wall time was approximately 20 minutes. The final combined loss was 0.0026. The calibrated parameter values are reported in Table 1.

All subsequent simulations—including all out-of-sample predictions reported in the Results—used this single calibrated parameter set with no further adjustment. The only quantities that change across simulations are scenario settings (M_i, protocol timing, number of trials) that are determined by each experiment's procedure, not by the modeler.

Simulation Protocols

Each experimental protocol was translated into a sequence of stimulus events specifying, for each trial, the active pathway, signal type (S^{ext}, S^{int}, or both), emotional arousal (M), branch coverage (η), and inter-trial and inter-session intervals. Between events, the system evolves with all signals set to zero, during which C cools, R decays, D dissolves according to its rate, and F remains unchanged (no signal implies no deformation and no repair).

For conditioning, each trial delivers S^{ext} to the fear pathway at the scenario's arousal level M_{cond}. For extinction, each trial delivers S^{ext} to the safety pathway at $M_{\text{ext}} = 1.0$. For retrieval trials, a single conditioned stimulus presentation simultaneously delivers S^{ext} (the tone as an external cue) and triggers S^{int} (internal reconstruction of the fear memory), with $\eta \approx 0.2$ reflecting the single-branch entry point. During extinction following retrieval, each extinction trial on the safety pathway also triggers automatic S^{int} on the fear pathway—the conditioned stimulus evokes partial fear reconstruction—but does not re-pump R on the fear pathway beyond its initial retrieval-induced level.

Specific protocol parameters (number of conditioning and extinction trials, inter-trial intervals, test delays) were matched to each published study as closely as the published methods sections permitted. Between conditioning and extinction sessions, a 1-day interval was imposed (overnight consolidation), during which C decays to near zero and D undergoes minor dissolution. Details of the numerical integration scheme are provided in Appendix A.

Competing Models

Two competing models were implemented for formal comparison.

Rescorla-Wagner

A revised Rescorla-Wagner model (Rescorla & Wagner, 1972) was implemented with separate associative strengths for fear (V_{fear}) and safety (V_{safety}), two learning rates (acquisition and extinction), an extinction-driven reduction rate on V_{fear}, independent

exponential decay rates for each associative strength, and a sigmoid temperature (6 free parameters total). This model was independently fitted to the Quirk spontaneous recovery data using the same optimizer.

Latent Cause Model

The latent cause model of (Gershman et al., 2017) was ported in full from the original implementation (available at https://github.com/sjgershm/memory-modification). This model combines a distance-dependent Chinese Restaurant Process for latent cause assignment with Rescorla-Wagner learning within each cause. Two versions were evaluated: (a) the model with default parameters as specified in the original publication, and (b) the model fitted to the Quirk spontaneous recovery data (7 free parameters, including a time-scale parameter mapping the model's abstract temporal units to days) using the same optimization procedure applied to the other models.

Both competing models were then tested on the same out-of-sample datasets as the Structural Crystallization framework, with no further parameter adjustment.

Transparency and Openness

No experimental data were collected for this study; all empirical benchmarks are previously published datasets cited in the text. All simulation code, calibration routines, and result files are available from the corresponding author upon request. The model and analysis were not preregistered. We report all calibration targets, all parameter values, all out-of-sample tests conducted, and all competing model comparisons; no analyses were excluded.

Results

Calibration: Spontaneous Recovery and Interval Gradient

Quirk (2002) Spontaneous Recovery

The model captured the negatively accelerating time course of spontaneous recovery following extinction (RMSE = 0.072; Figure 1a). Predicted percent rebound increased

monotonically from 0.43 at day 1 to 0.99 at day 14, tracking the observed trajectory from 0.38 to 1.00. The largest residual occurred at day 2 (model: 0.51, observed: 0.63), reflecting a slight underestimation of early recovery. The overall shape—rapid initial recovery decelerating toward asymptote—was well captured.

The mechanism producing this curve is the differential dissolution of competing pathways. The fear pathway, formed at $M_i = 5.58$, has a half-life of approximately 34 days. The safety pathway, formed during extinction at $M_i = 1.0$, has a substantially shorter half-life. Spontaneous recovery reflects the progressive loss of the safety pathway's competitive suppression as it dissolves faster than the fear pathway, not a strengthening of the fear memory itself. This mechanism depends on the M_i-dependent dissolution asymmetry in Equation 4—a feature that no single-strength model possesses.

Monfils (2009) Interval Gradient

The model produced a monotonically increasing gradient of spontaneous recovery as a function of the retrieval-to-extinction interval (Figure 1b): 10 min (0.001) $<$ 1 hr (0.004) $<$ 6 hr (0.198) $<$ 24 hr (0.760). All hard constraints were satisfied: the gradient was monotonic, the spread exceeded 0.15, and the 10-minute condition produced recovery well below 0.5.

The calibrated reconsolidation window time constant was $\tau_R = 0.168$ days (approximately 4 hours). This value determines the gradient's shape: at short intervals, R remains elevated during extinction and each extinction trial drives fidelity loss on the fear pathway; at long intervals, R has decayed to near zero and extinction proceeds without deforming the fear trace. The transition between these regimes is continuous, governed by the exponential decay of R. This gradient arises from the interaction between labilization (R) and fidelity loss (F)—two variables that exist only because accumulation and fidelity are modeled separately.

Out-of-Sample Predictions: Fear Conditioning

The following predictions were generated from the calibrated parameter set with no further adjustment. The only quantities that change are scenario settings determined by each experiment's published procedure.

Retrieval-Extinction Blocking (Schiller et al., 2010)

Between-group. The model predicted that retrieval-extinction would block spontaneous recovery, consistent with the observed data (Figure 2a). For standard extinction, the model predicted day-1 recovery of 0.43; for retrieval-extinction, the model predicted 0.00. The observed values were 0.57 and 0.00, respectively.

The mechanism is fidelity loss accumulation. During retrieval-extinction at a 10-minute interval, R remains elevated throughout extinction. Each extinction trial triggers automatic internal reconstruction of the fear memory ($S^{\text{int}}_{\text{fear}} > 0$), which passes through the labile fear structure at low coverage ($\eta \approx 0.2$). This drives F_{fear} to 0.73, reducing effective fear strength from $D_{\text{fear}} = 0.93$ to $D_{\text{fear}}(1 - F_{\text{fear}}) = 0.25$. This effective strength falls below D_{safety} and remains below it for months, preventing recovery. Standard extinction, by contrast, never opens a reconsolidation window ($R = 0$ throughout), so F_{fear} remains at zero and recovery depends entirely on the dissolution race between fear and safety pathways.

This result illustrates why the D/F separation is necessary: standard extinction and retrieval-extinction produce the same D_{safety} increase, but only retrieval-extinction additionally modifies F_{fear}. A single-strength model, which cannot distinguish these two routes to reduced fear expression, predicts identical outcomes for both procedures.

Within-subject. The model predicted lower fear expression for the reminded stimulus (CSa) than the non-reminded stimulus (CSb) at all clinically relevant time points (Figure 2b): day 0 (0.00 vs. 0.44), day 1 (0.00 vs. 0.52), day 7 (0.00 vs. 0.88), and day 30 (0.02 vs. 1.00). This dissociation arises because only CSa underwent retrieval before extinction, so only the CSa fear pathway accumulated fidelity loss. The CSb fear pathway, never retrieved, retained $F = 0$ and followed the standard spontaneous recovery trajectory.

At 365 days, the model predicted convergence (0.99 vs. 1.00), reflecting the eventual dissolution of D_{safety} to near zero. Even the reduced effective fear strength of CSa ($D_{\text{eff}} = 0.25$) is sufficient to drive fear expression through the sigmoid readout once the competing safety pathway has fully dissolved. This is somewhat faster than the >1-year persistence reported by Schiller et al. (2010), suggesting that the model underestimates the long-term stability of fidelity loss or that residual safety memory persists longer than predicted.

Stress-Enhanced Fear Learning (Rau & Fanselow, 2009)

The model was applied to the stress-enhanced fear learning (SEFL) paradigm of Rau and Fanselow (2009), in which groups of rats received 0, 1, 4, or 15 prior shocks before a single conditioning trial in a novel context.

Day 1: Saturation. At day 1, the model predicted freezing levels of 0.05 (0 shocks), 0.29 (1 shock), 0.31 (4 shocks), and 0.31 (15 shocks). The model captured the saturation between 4 and 15 shocks (both ≈ 0.31) but underestimated the step from 1 to 4 shocks observed in the data (0.21 vs. 0.65). This day-1 underestimation arises because the single conditioning trial is gate-limited: regardless of M, one trial can only push D through the sigmoid gate once, producing similar D values across high-M groups.

Long-term divergence: The SEFL effect. The framework's distinctive prediction for SEFL is not about initial acquisition but about long-term persistence (Figure 3). The prior shock history determines M_i, which locks the fear memory's dissolution resistance. Half-lives span four orders of magnitude: 4.7 days (0 shocks, $M_i = 1.0$), 259 days (1 shock, $M_i = 6.4$), 117,455 days (4 shocks, $M_i = 15.0$), and 3,850,429 days (15 shocks, $M_i = 19.9$).

At day 1, all high-shock groups show similar D (≈ 0.30). By day 30, the 0-shock group has dissolved to 0.001 while the 4- and 15-shock groups remain at 0.305. By day 365, the divergence is complete: the 0-shock memory is gone, the 1-shock memory has declined to 0.11, and the 4- and 15-shock memories are unchanged.

The framework thus reinterprets SEFL: prior stress does not make animals "scared deeper" (initial D is similar across high-M groups) but "scared harder"—M_i locks the memory into a dissolution-resistant state. The behavioral consequence is not visible at short delays but emerges dramatically over time. This reinterpretation depends on the dissolution term in Equation 4, where M_i enters the exponential suppression factor—a mechanism unavailable to models that lack arousal-dependent decay.

Appetitive Extinction Durability (Woods & Bouton, 2008)

The model was applied to appetitive conditioning ($M = 1.0$, $M_i = 1.0$) following the appetitive protocol of Woods and Bouton (2008) (40 conditioning trials, 20 extinction trials), with only the arousal level changed to $M = 1.0$.

The model predicted near-flat appetitive recovery: 0.53 at day 0, 0.53 at day 1, 0.54 at day 4, and 0.54 at day 14 (Figure 4). In contrast, fear recovery over the same period was dramatic: 0.35 at day 0, escalating to 0.99 at day 14.

The mechanism is symmetric dissolution. For fear conditioning, $M_i = 5.58$ for the fear pathway versus $M_i = 1.0$ for the safety pathway creates an asymmetry: safety dissolves while fear persists, producing progressive recovery. For appetitive conditioning, both the appetitive pathway and the safety pathway form at $M_i = 1.0$. Both dissolve at the same rate. The competitive balance established during extinction is maintained indefinitely because neither pathway has a dissolution advantage.

In one sentence: fear relapses because fear is hard and therapy is soft; appetitive conditioning does not relapse because both sides are equally soft and age together. This prediction—which follows from a single parameter (M_i) with no additional mechanism—is consistent with the well-established finding that appetitive extinction is more durable than fear extinction (Rescorla, 2004; Woods & Bouton, 2008).

Competing Model Comparison

Rescorla-Wagner

The revised Rescorla-Wagner model achieved a lower Quirk RMSE (0.059) than the Structural Crystallization framework (0.072), reflecting the additional flexibility of its 6 free parameters applied to a single dataset. However, when tested out-of-sample, the Rescorla-Wagner model failed on both critical tests: it did not block spontaneous recovery under retrieval-extinction (predicted recovery: 0.42 for standard extinction vs. 0.43 for retrieval-extinction) and produced a flat Monfils gradient (all conditions ≈ 0.97). These failures are structural, not parametric: the Rescorla-Wagner framework contains no reconsolidation mechanism, so retrieval before extinction has no differential effect.

Latent Cause Model

With default parameters (not fitted to any dataset), the latent cause model of Gershman et al. (2017) successfully predicted Schiller-style blocking (0.07 vs. 0.99), a Monfils-style gradient (short intervals ≈ 0.07, long intervals $= 0.85$), and the CSa $<$ CSb within-subject dissociation (0.05 vs. 0.31). These successes reflect the model's core mechanism: retrieval before extinction reactivates the original cause, allowing extinction to modify it directly rather than creating a new competing cause.

However, when the latent cause model was fitted to the Quirk spontaneous recovery data (RMSE $= 0.050$, the best fit among all models), its out-of-sample predictions collapsed. The Schiller blocking effect disappeared (retrieval-extinction: 0.35 vs. standard: 0.42), and the Monfils gradient inverted (short intervals produced *higher* recovery than long intervals). The time-scale parameter required to reproduce the multi-day Quirk recovery curve disrupted the cause-switching dynamics responsible for the model's retrieval-extinction predictions.

Summary

Table 2 and Figure 5 summarize the comparison. The Structural Crystallization framework is the only model that simultaneously fits the calibration data and correctly predicts out-of-sample phenomena with the same parameter set. The Rescorla-Wagner model fits calibration data well but lacks the mechanistic structure for reconsolidation phenomena. The latent cause model captures reconsolidation phenomena with default parameters but cannot maintain these predictions when constrained to fit empirical time courses.

An additional qualitative difference concerns the predicted shape of the interval gradient. The Structural Crystallization framework predicts a continuous gradient governed by the exponential decay of R with time constant τ_R, producing a smooth transition from low to high recovery as the interval increases. The latent cause model (with default parameters) predicts a step function: cause-switching is discrete, producing all-or-nothing transitions. The available Monfils data, though limited in resolution, appear more consistent with a continuous gradient.

Extensions to Human Declarative Memory

To assess whether the framework generalizes beyond fear conditioning, it was applied to four benchmark phenomena from the human declarative memory literature. All simulations used neutral encoding parameters ($M = 1.0$, $M_i = 1.0$) with the calibrated parameter set; no parameters were re-fitted.

Testing Effect

The testing effect refers to the finding that retrieval practice produces better long-term retention than restudying, despite restudying producing superior performance at short delays—a crossover interaction (Karpicke & Roediger, 2008; Roediger & Karpicke, 2006). The framework was applied by modeling study trials as referenced access (S^{ext}) and test trials as unreferenced access (S^{int}), following the study–study (SS) versus study–test (ST) design of Roediger and Karpicke (2006), Experiment 1.

The model produced the correct crossover pattern: at 5 minutes, SS (1.000) exceeded ST (1.000, marginal advantage); at 2 days, ST (0.960) exceeded SS (0.948); at 1 week, ST (0.850) exceeded SS (0.804). The direction of the crossover matched the observed data (SS: 0.81 $\rightarrow$ 0.54 $\rightarrow$ 0.42; ST: 0.75 $\rightarrow$ 0.68 $\rightarrow$ 0.56) at all three delays. The mechanism is that test trials, as unreferenced access, push both C and R; the additional C boost from retrieval effort produces slightly higher D advancement, which becomes visible only after the short-term C component has decayed. The absolute magnitude of the crossover was smaller in the model ($\sim$5%) than in the data ($\sim$14%), likely reflecting the fact that the parameters were calibrated to high-arousal fear conditioning rather than neutral verbal learning.

This result demonstrates that the framework's distinction between referenced and unreferenced access has consequences beyond reconsolidation: it naturally produces the testing effect crossover as a zero-parameter prediction in a domain (human declarative memory) far removed from the calibration data.

Spacing Effect

The spacing effect—the advantage of distributed over massed practice (Cepeda et al., 2006)—was not successfully captured. The model predicted massed D (0.58) $>$ spaced D (0.22), the opposite of the observed pattern. This failure arises because $\alpha_{C0} = 191.5$ (calibrated to fear conditioning, where a single trial must drive rapid crystallization) overwhelms the $(1 - C)$ bandwidth limitation that is the framework's mechanism for the spacing effect. The $(1 - C)$ factor does produce diminishing returns for massed trials, but at the calibrated α_{C0} the effect is too weak to overcome the raw advantage of continuous stimulation. This result indicates that α_{C0} likely needs to scale with task context—a theoretically meaningful constraint suggesting that the effective push rate differs between high-arousal single-trial paradigms and low-arousal multi-trial paradigms.

Retrieval-Induced Forgetting

Retrieval-induced forgetting (RIF) was partially captured. Following the category-exemplar paradigm of Anderson et al. (1994), the model predicted D_{eff} for practiced items (Rp+) of 0.373, exceeding both non-practiced items from practiced categories (Rp−; 0.324) and baseline items (Nrp; 0.324). The Rp+ > Nrp advantage (+15%) matched the observed direction (73.6% vs. 48.4%). However, the model failed to produce Rp− < Nrp: both were 0.324, whereas the observed data showed Rp− (37.5%) substantially below Nrp (48.4%).

The framework captures retrieval enhancement (Rp+ benefits from additional C and D advancement during practice) but not retrieval-induced suppression (Rp− should be actively inhibited by practice of competing items). The inter-pathway interference term $\gamma \cdot h_i \cdot C_i$ is too weak at current parameter values to produce measurable suppression. This result aligns the framework with competition-based accounts of RIF (Anderson, 2003; Anderson et al., 1994): the framework naturally produces facilitation of practiced items without active suppression of competitors. Whether RIF requires a dedicated inhibitory mechanism or can be captured by strengthening the interference term is an open theoretical question.

Flashbulb Memory

The framework provides a natural account of flashbulb memories—vivid, confident, but often inaccurate recollections of emotionally significant events. Under high M at encoding, D is driven to a high value (the memory feels vivid and accessible) while subsequent unreferenced retrieval (rumination, retelling) progressively increases F (the memory's structure drifts from its original form). The result is a memory that is simultaneously high in D (perceived vividness) and high in F (objective inaccuracy). This pattern matches the empirical dissociation reported by Talarico and Rubin (2003): flashbulb memories showed declining consistency over time (increasing F) while subjective vividness ratings remained elevated (reflecting high D). This account is qualitative rather

than quantitative, but it illustrates the D/F separation's ability to explain phenomena where subjective experience and objective accuracy diverge.

Exploratory: Ebbinghaus Forgetting Curve

As an exploratory test, the model was applied to a neutral declarative memory scenario ($M = 1.0$, $M_i = 1.0$, 8 S^{ext} encoding trials) with no parameter changes.

The predicted retention curve exhibited two natural timescales (Figure 6): rapid early decay driven by C cooling (time constant $\approx$ 82 minutes) and slower long-term decay driven by D dissolution (half-life $\approx$ 22 days). A power-law fit to the predicted D-only curve yielded an exponent of $b = 0.24$; the combined $D + C$ curve yielded $b = 0.31$. Both fall within the empirically observed range of 0.06–0.70 for forgetting curve exponents (Rubin & Wenzel, 1996; Wixted, 2004).

The quantitative fit to the classical Ebbinghaus data was limited—the model overestimates retention at short delays and underestimates it at long delays—reflecting the fact that the framework was calibrated to fear conditioning protocols with very different temporal dynamics. This result is presented as a supplementary exploration demonstrating that the framework's dissolution mechanics naturally produce power-law-like forgetting without additional assumptions, not as a formal quantitative prediction.

General Discussion

The central claim of this article is that memory is not scalar but composite: it consists of at least two independent dimensions—structural existence and representational integrity—which are differentially modified by retrieval. This is not a claim about a particular model. It is a claim about the minimum representational complexity required for any adequate theory of memory. A framework that collapses these dimensions into a single quantity is subject to a representational constraint that prevents it from simultaneously accounting for why memories persist, why they change, and why the same act of retrieval can produce either outcome depending on conditions.

The Structural Crystallization framework serves as a formal proof of this claim's

computational sufficiency. Calibrated once to two datasets, the framework correctly predicted—without parameter adjustment—outcomes across five independent benchmarks spanning fear conditioning, stress-enhanced learning, cross-valence extinction, and human declarative memory. It also produced partial accounts of retrieval-induced forgetting and flashbulb memory, while failing on the spacing effect. Competing models, which represent memory as a single quantity, could not simultaneously fit calibration data and predict these phenomena with the same parameters.

The pattern of successes and failures is itself informative. The framework succeeds precisely where the composite nature of memory is load-bearing—phenomena that require distinguishing structural existence from representational integrity—and fails where other factors (task-specific encoding rates) dominate. This pattern suggests that the existence–integrity separation is not merely one possible modeling choice but a necessary structural feature of any framework that aspires to unified coverage of memory persistence and modification.

Theoretical Contributions

Structural Existence and Representational Integrity as Independent Dimensions

The framework's central theoretical contribution is the demonstration that memory requires at least two independent dimensions: one tracking structural existence (D, how much of the trace has been consolidated) and one tracking representational integrity (F, how faithfully the existing structure preserves its original encoding). Existing models conflate these into a single quantity—associative strength in Rescorla-Wagner, posterior probability in latent cause models. The separation makes a qualitative difference: it allows a memory to retain its full structural extent while becoming progressively distorted, producing a reduction in behavioral expression that is mechanistically distinct from forgetting.

This distinction resolves an ambiguity that pervades the memory modification

literature. When a retrieval-extinction procedure reduces fear expression, existing frameworks must choose between two interpretations: the fear memory was weakened (less of it exists) or a competing safety memory was strengthened (the fear memory is unchanged but suppressed). The D/F separation provides a third possibility: the fear memory retains its full structural extent (D unchanged) but has been structurally deformed (F increased). This third interpretation naturally explains why retrieval-extinction effects resist spontaneous recovery, renewal, and reinstatement—the deformation is a property of the fear trace itself, not of a competing trace that might dissolve or lose context-dependence.

In the framework, spontaneous recovery after standard extinction reflects the dissolution of a competing safety trace (D_{safety} decays), whereas the persistent blocking of recovery after retrieval-extinction reflects structural deformation of the fear trace itself (F_{fear} increases). These are different processes operating on different variables, yet both emerge from the same equation system. This mechanistic separation would be impossible in any framework with a single strength variable.

The Asymmetry Between Access Modes

The distinction between referenced and unreferenced access provides a principled account of why retrieval can either strengthen or weaken a memory depending on the conditions. Referenced access (high branch coverage, multi-constraint convergence) is structurally protective: it advances crystallization and corrects prior deformation without inducing labilization. Unreferenced access (low branch coverage, few-constraint reconstruction) is structurally risky: it induces labilization, and any signal passing through the labile structure at low coverage drives deformation.

This asymmetry is not a post hoc distinction added to explain reconsolidation data. It follows from the physical metaphor of multi-branch crystallization, where the number of constraints determines whether activation is controlled or uncontrolled. A crystal being annealed under uniform pressure from all directions (referenced access) will settle into a more ordered state. A crystal being stressed from a single point while the rest of its

structure is unsupported (unreferenced access) may deform at the point of stress while strengthening locally. The metaphor generates the asymmetry; the equations formalize it.

This account connects to the broader literature on the paradoxical effects of retrieval. Retrieval practice enhances long-term retention (the testing effect; Roediger and Karpicke, 2006), yet retrieval can also induce forgetting of competing items (retrieval-induced forgetting; Anderson et al., 1994) and open memories to modification (reconsolidation; Nader et al., 2000). These seemingly contradictory effects are unified under the access-mode distinction: the consequences of retrieval depend on the constraint density (η) under which the memory is accessed.

Emotional Arousal as a Permanence Lock

The parameter M_i—emotional arousal at encoding, locked at formation—plays a dual role that unifies several otherwise disparate phenomena. It accelerates initial crystallization (explaining rapid fear acquisition) and exponentially suppresses subsequent dissolution (explaining the extraordinary persistence of traumatic memories). The same mechanism, with no additional assumptions, accounts for three phenomena that are typically explained by separate theories:

1. *Stress-enhanced fear learning*: prior shocks set M_i, producing memories with half-lives spanning four orders of magnitude—not "scared deeper" (initial D is similar) but "scared harder" (dissolution resistance differs by factors of 10^4).
2. *Fear extinction fragility*: the M_i asymmetry between fear ($M_i = 5.58$) and safety ($M_i = 1.0$) pathways drives differential dissolution, making safety traces dissolve while fear traces persist.
3. *Appetitive extinction durability*: both appetitive and safety traces form at $M_i = 1.0$, so both dissolve at the same rate. The competitive balance is maintained indefinitely.

The M_i lock provides a parsimonious account of why emotional memories persist: not because they are stored in a special system or processed by a dedicated circuit, but because the encoding conditions permanently alter a single parameter that controls the

trace's resistance to a universal dissolution process. This is consistent with the extensive evidence that emotional arousal at encoding enhances memory consolidation through neuromodulatory mechanisms (McGaugh, 2000), but reframes the effect as a dissolution-resistance parameter rather than a consolidation-enhancement parameter. The distinction matters: enhancement of consolidation predicts that emotional memories start stronger, whereas dissolution resistance predicts that emotional memories *last* longer—a prediction more consistent with the SEFL data, where initial acquisition is similar across stress levels but long-term persistence differs dramatically.

Implications for Existing Theoretical Debates

The results of the model comparison and out-of-sample predictions speak to several active debates in the memory literature.

Erasure Versus New Learning in Extinction

The longstanding debate over whether extinction erases the original conditioning or creates new, competing learning (Bouton, 2004; Rescorla, 2004) has been complicated by retrieval-extinction findings that suggest a third possibility: modification of the original trace without erasure. The framework resolves this three-way ambiguity. Standard extinction is new learning (D_{safety} increases, D_{fear} and F_{fear} unchanged)—fully consistent with Bouton's account. Retrieval-extinction is modification without erasure (D_{fear} unchanged, F_{fear} increases)—the fear trace still exists at full structural extent but has been deformed. The distinction between these outcomes is not in the extinction procedure itself but in whether the fear trace was labilized before extinction began. The framework thus provides a unified account in which both "new learning" and "modification" are correct descriptions of extinction—but of different extinction protocols operating through different mechanisms within the same equation system.

Competition Versus Inhibition in Retrieval-Induced Forgetting

The framework's partial capture of retrieval-induced forgetting positions it on the competition side of the competition-versus-inhibition debate (Anderson, 2003; Anderson

et al., 1994). The model naturally produces facilitation of practiced items (Rp+ benefits from additional $C \rightarrow D$ advancement) but does not produce active suppression of unpracticed competitors (Rp− = Nrp). The inter-pathway interference term $\gamma \cdot h_i \cdot C_i$ is too weak at current parameter values to generate measurable suppression.

This result is diagnostic rather than merely a limitation. It reveals which aspects of RIF the accumulation–fidelity framework can and cannot capture. The competition component—practiced items are strengthened relative to baseline—falls naturally out of the crystallization dynamics. The inhibition component—unpracticed items from practiced categories are suppressed below baseline—would require either a stronger interference mechanism or a dedicated inhibitory process not currently in the framework. Whether RIF ultimately requires an inhibitory mechanism beyond resource competition remains an open question, but the framework cleanly delineates where the boundary lies.

The Temporal Structure of Reconsolidation

The framework makes a specific prediction about the shape of the reconsolidation window that differs from the latent cause model. The Structural Crystallization framework predicts a *continuous* gradient: as the retrieval-to-extinction interval increases, R decays exponentially, and the effectiveness of extinction at modifying the fear trace decreases smoothly. The latent cause model (with default parameters) predicts a *step function*: cause-switching is a discrete event that either occurs or does not.

The available data from Monfils et al. (2009) are broadly consistent with a continuous gradient but do not have sufficient temporal resolution to definitively distinguish between these predictions. High-resolution interval studies—testing at 2, 3, 5, and 8 hours post-retrieval—would provide a critical test. The framework predicts that the transition from full blocking to no blocking should follow a smooth exponential with $\tau_R \approx 4$ hours, not an abrupt threshold.

Clinical Implications

The framework offers a mechanistic account of why exposure therapy works when it works, and why it fails when it fails.

Standard exposure therapy operates through referenced access (S^{ext}) to the safety pathway: each extinction trial builds D_{safety} while leaving the fear trace intact ($F_{\text{fear}} = 0$). The therapeutic effect depends entirely on the safety trace's ability to outcompete the fear trace. Because the fear trace was formed at high M_i and the safety trace at low M_i, the safety trace dissolves faster—producing the relapse that is the central clinical challenge of exposure-based treatments (Craske et al., 2008, 2014).

Retrieval-extinction operates through a fundamentally different mechanism: by opening a reconsolidation window ($R > 0$) and then passing extinction signals through the labile fear trace at low coverage, it accumulates fidelity loss on the fear trace itself. The therapeutic effect is not suppression by a competing trace but structural modification of the pathological trace. This mechanism is inherently more durable because F does not spontaneously recover and is not context-dependent.

The framework predicts that the effectiveness of retrieval-extinction should depend on three factors: (a) whether the retrieval event successfully induces labilization ($R > 0$, requiring unreferenced access with low η), (b) the interval between retrieval and extinction (shorter is better, governed by $\tau_R \approx 4$ hours), and (c) the number of extinction trials during the window (more trials accumulate more F). These predictions align with the emerging clinical literature on reconsolidation-based interventions and provide specific quantitative constraints for optimizing treatment protocols.

An additional clinical implication concerns the persistence of trauma memories. The framework explains PTSD-like persistence not through a special pathological mechanism but through the normal operation of the M_i lock: extreme emotional arousal at encoding produces a memory with a dissolution half-life of years to decades, making it effectively permanent under normal conditions. This account predicts that any intervention targeting

the fear trace must operate through fidelity loss (F) rather than dissolution (D), because dissolution is exponentially suppressed at high M_i. In practical terms, helping a trauma patient "forget" the event through passive time or distraction is predicted to be ineffective; modifying the trace through controlled reactivation and reconsolidation-based procedures is predicted to be the only viable route.

Boundary Conditions and Limitations

The framework's failures are as informative as its successes, because they delineate the conditions under which the accumulation–fidelity distinction is load-bearing versus conditions where other factors dominate.

The Spacing Effect as a Diagnostic Failure

The spacing effect failure—the model predicted massed > spaced, the opposite of the observed pattern—is not a random miss. It has a precise diagnosis: the crystallization push rate α_{C0}, calibrated to high-arousal single-trial fear conditioning where a single trial must produce substantial C advancement, overwhelms the $(1 - C)$ bandwidth limitation that is the framework's mechanism for the spacing effect. The $(1 - C)$ factor does produce diminishing returns for massed trials, but at $\alpha_{C0} = 191.5$ the effect is too weak to overcome the raw advantage of continuous stimulation.

This failure diagnoses a parameter transfer boundary: α_{C0} likely needs to scale with task context. In fear conditioning, a single high-arousal trial must drive the system through the $C \to D$ gate in one pulse. In neutral multi-trial learning, the push rate per trial should be much lower, allowing the $(1 - C)$ bandwidth limit to operate as intended. The prediction that α_{C0} is context-dependent—specifically, that it should be lower for low-arousal, multi-trial paradigms than for high-arousal, single-trial paradigms—is itself an empirically testable claim about the boundary conditions of the framework.

Parameter Count and Identifiability

The model contains 10 free parameters calibrated to two datasets. Although the sensitivity analysis revealed clean separation into two functionally distinct parameter

groups (crystallization–dissolution: D_c, α_{C0}, δ_C, β_0; deformation: ϕ, ψ, τ_R) with zero cross-group sensitivity, the parameter count is not trivial.

Three parameter pairs exhibited near-perfect correlations: β_0 and D_c ($r = .999$), α_{C0} and κ ($r = .995$), and D_c and M_{cond} ($r = .987$). These correlations are structural, not fitting artifacts: in each case, the two parameters enter the equations as a product or ratio ($\beta_0 \cdot e^{-D \cdot M / D_c}$ renders β_0 and D_c exchangeable in their effect on dissolution rate). The identifiable quantities are the composite ratios β_0 / D_c and $\alpha_{C0} \cdot \kappa$, not the individual parameters. Future work should explore whether structural constraints—fixing β_0 / D_c as a single parameter, for example—can reduce the effective dimensionality without sacrificing predictive accuracy.

Branch Coverage as a Binary Setting

The framework treats branch coverage η as a binary scenario setting: high ($\sim$0.9) for referenced access, low ($\sim$0.1–0.2) for unreferenced access. In reality, coverage likely varies continuously across retrieval conditions—from a rich multi-modal cue (high η) to a single partial cue (low η) to completely spontaneous recall (very low η). A more granular treatment of η, potentially linked to the number and quality of available cues, would extend the framework's applicability but at the cost of additional complexity. The binary treatment is sufficient for the phenomena addressed here because the relevant experiments involve clear manipulations of access mode (conditioning trials vs. retrieval trials), but it would need refinement for phenomena where coverage varies more subtly.

Fidelity Loss as a Scalar

Fidelity loss F is modeled as a scalar, capturing how much deformation has occurred but not in which direction. The current framework cannot distinguish between deformation that reduces fear expression (therapeutic, as in retrieval-extinction) and deformation that distorts the memory in maladaptive ways (as in false memory formation or rumination-driven distortion in PTSD). Extending F to a vector quantity with directional information is a natural next step but would substantially increase the framework's

complexity. For the phenomena addressed here—where the behavioral readout is a scalar competition between fear and safety pathways—the scalar treatment is sufficient.

Long-Term Stability of Fidelity Loss

The model's prediction for the Schiller 1-year follow-up was somewhat faster than observed: the model predicted near-convergence of reminded and non-reminded stimuli at 365 days, whereas the empirical data showed persistent differentiation beyond one year (Schiller et al., 2010). This suggests that either the fidelity loss mechanism is more stable than modeled (perhaps F should have an even weaker natural recovery term, or none at all) or that residual safety memory persists longer than the dissolution equation predicts. Both possibilities are testable.

Empirical Predictions

The framework generates several testable predictions that go beyond the phenomena reported here.

Prediction 1: The retrieval-extinction interval gradient is continuous, not discrete. The framework predicts a smooth exponential transition from full blocking to no blocking as the retrieval-to-extinction interval increases, governed by $\tau_R \approx 4$ hours. The latent cause model predicts a step function. High-resolution interval studies (e.g., 2 hr, 3 hr, 5 hr, 8 hr) could distinguish between these predictions.

Prediction 2: SEFL divergence emerges over time, not at acquisition. The framework predicts that groups receiving different numbers of prior shocks should show similar fear levels shortly after conditioning but diverge dramatically over weeks, with the divergence pattern determined by M_i-dependent half-lives. Longitudinal SEFL studies with multiple post-conditioning test points could test this prediction.

Prediction 3: Appetitive reconsolidation should be more effective than fear reconsolidation. Because appetitive memories have lower M_i and thus lower D, the effective strength reduction produced by a given amount of fidelity loss should be proportionally larger for appetitive than for fear memories. Retrieval-extinction procedures applied to

appetitive conditioning should produce more complete and more durable blocking of spontaneous recovery than the same procedure applied to fear conditioning.

Prediction 4: Multiple retrieval events without intervening extinction should accumulate fidelity loss. Each unreferenced retrieval drives R upward, and if subsequent signals pass through the labile structure, F increases. Repeated retrieval-only trials (without extinction) should produce measurable fidelity loss on the retrieved memory, observable as reduced fear expression even without extinction training—but only if inter-retrieval intervals are short enough to maintain elevated R.

Prediction 5: The crystallization push rate should scale with task arousal. The spacing effect failure predicts that α_{C0} is lower for neutral, multi-trial learning paradigms than for emotional, single-trial conditioning. This could be tested by measuring the spacing effect across a range of arousal levels: the framework predicts that the spacing advantage should emerge as arousal (and therefore α_{C0}) decreases, because lower α_{C0} allows the $(1 - C)$ bandwidth limit to produce meaningful diminishing returns for massed practice.

Prediction 6: Fidelity loss should be detectable independently of accumulation. If D and F are truly independent dimensions, it should be possible to find conditions under which D remains constant but F changes, or vice versa. The framework predicts that repeated unreferenced retrieval at short intervals should increase F without substantially changing D (because R-driven deformation does not affect the D equation). Conversely, passive time should decrease D without changing F (because dissolution operates on D but F requires active signal to change). Memory measures that distinguish structural extent from content accuracy—such as comparing confidence ratings (D-dependent) with consistency checks (F-dependent)—could provide converging evidence.

Conclusion

The question that motivated this work was not "Can we build a model that fits more data?" but "What is the minimum representational complexity that any theory of memory must have?" The answer we propose is that memory is composite, not scalar: it

requires at least two independent dimensions—structural existence and representational integrity—which are differentially modified by retrieval. This is not a modeling convenience but a representational necessity. Any framework that collapses these dimensions into a single quantity inherits a structural constraint that prevents it from explaining why extinction is temporary but retrieval-extinction is lasting, why trauma memories persist for decades, or why the same act of remembering can either preserve or destroy what is remembered. The Structural Crystallization framework serves as a formal proof that two dimensions, combined with a distinction between protective and risky retrieval, are computationally sufficient to generate the empirical record. The contribution is the theoretical claim; the model is its instantiation.

References

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, *49*(4), 415–445. https://doi.org/10.1016/j.jml.2003.08.006

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin & Review*, *1*(4), 522–536. https://doi.org/10.3758/BF03210948

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (pp. 35–67, Vol. 2). Erlbaum.

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological Bulletin*, *114*(1), 80–99. https://doi.org/10.1037/0033-2909.114.1.80

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning & Memory*, *11*(5), 485–494. https://doi.org/10.1101/lm.78804

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Craske, M. G., Kircanski, K., Zelikowsky, M., Mystkowski, J., Chowdhury, N., & Baker, A. (2008). Optimizing inhibitory learning during exposure therapy. *Behaviour Research and Therapy*, *46*(1), 5–27. https://doi.org/10.1016/j.brat.2007.10.003

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., & Vervliet, B. (2014). Maximizing exposure therapy: An inhibitory learning approach. *Behaviour Research and Therapy*, *58*, 10–23. https://doi.org/10.1016/j.brat.2014.04.006

Gershman, S. J., Monfils, M.-H., Norman, K. A., & Niv, Y. (2017). The computational nature of memory modification (M. J. Frank, Ed.). *eLife*, *6*, e23763. https://doi.org/10.7554/eLife.23763

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Lee, J. L. C., Nader, K., & Schiller, D. (2017). An Update on Memory Reconsolidation Updating. *Trends in Cognitive Sciences*, *21*(7), 531–545. https://doi.org/10.1016/j.tics.2017.04.006

McGaugh, J. L. (2000). Memory–a Century of Consolidation. *Science*, *287*(5451), 248–251. https://doi.org/10.1126/science.287.5451.248

Monfils, M.-H., Cowansage, K. K., Klann, E., & LeDoux, J. E. (2009). Extinction-reconsolidation boundaries: Key to persistent attenuation of fear memories. *Science*, *324*(5929), 951–955. https://doi.org/10.1126/science.1167975

Nader, K., Schafe, G. E., & Le Doux, J. E. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, *406*(6797), 722–726. https://doi.org/10.1038/35021052

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552. https://doi.org/10.1037/0033-295X.87.6.532

Quirk, G. J. (2002). Memory for Extinction of Conditioned Fear Is Long-lasting and Persists Following Spontaneous Recovery. *Learning & Memory*, *9*(6), 402–407. https://doi.org/10.1101/lm.49602

Rau, V., & Fanselow, M. S. (2009). Exposure to a stressor produces a long lasting enhancement of fear learning in rats. *Stress*, *12*(2), 125–133. https://doi.org/10.1080/10253890802137320

Rescorla, R. A. (2004). Spontaneous recovery. *Learning & Memory*, *11*(5), 501–509. https://doi.org/10.1101/lm.77504

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*(4), 734–760. https://doi.org/10.1037/0033-295X.103.4.734

Schiller, D., Monfils, M.-H., Raio, C. M., Johnson, D. C., LeDoux, J. E., & Phelps, E. A. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature*, *463*(7277), 49–53. https://doi.org/10.1038/nature08637

Storn, R., & Price, K. (1997). Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization*, *11*(4), 341–359. https://doi.org/10.1023/A:1008202821328

Talarico, J. M., & Rubin, D. C. (2003). Confidence, not consistency, characterizes flashbulb memories. *Psychological Science*, *14*(5), 455–461. https://doi.org/10.1111/1467-9280.02453

Wixted, J. T. (2004). The Psychology and Neuroscience of Forgetting. *Annual Review of Psychology*, *55*, 235–269. https://doi.org/10.1146/annurev.psych.55.090902.141555

Woods, A. M., & Bouton, M. E. (2008). Immediate extinction causes a less durable loss of performance than delayed extinction following either fear or appetitive conditioning. *Learning & Memory*, *15*(12), 909–920. https://doi.org/10.1101/lm.1078508

Table 1

Model parameters, organized by type.

Type	Parameter	Value	Meaning
Fixed	C_{thresh}	0.80	Branch completion threshold
	w	0.05	Sigmoid gate sharpness
	ρ	1.0	Labilization injection strength
Fitted	α_{C0}	191.46	Base crystallization push rate
	δ_C	17.55	Branch cooling rate (day^{-1})
	κ	16.09	Trunk advance rate
	β_0	0.168	Base dissolution rate (day^{-1})
	D_c	0.428	Crystallization threshold
	τ	0.0547	Readout sigmoid temperature
	ϕ	9.09	Deformation rate
	ψ	15.10	Repair rate
	τ_R	0.168 days	Reconsolidation window time constant
	M_{cond}	5.58	Conditioning arousal (scenario)
Scenario	M_{ext}	1.0	Extinction arousal
	B	7.0	Global resource budget
	γ	0.1	Inter-pathway interference

Table 2

Model comparison across calibration and out-of-sample tests.

Test	Structural Crystallization	Rescorla-Wagner	Latent Cause (default)	Latent Cause (fitted)
Quirk RMSE	0.072	**0.059**	—	**0.050**
Schiller blocking	**Yes**	No	Yes	No
Monfils gradient	**Yes**	No	Yes (step)	Inverted
SEFL persistence	**Yes**	—	—	—
Appetitive durability	**Yes**	—	—	—

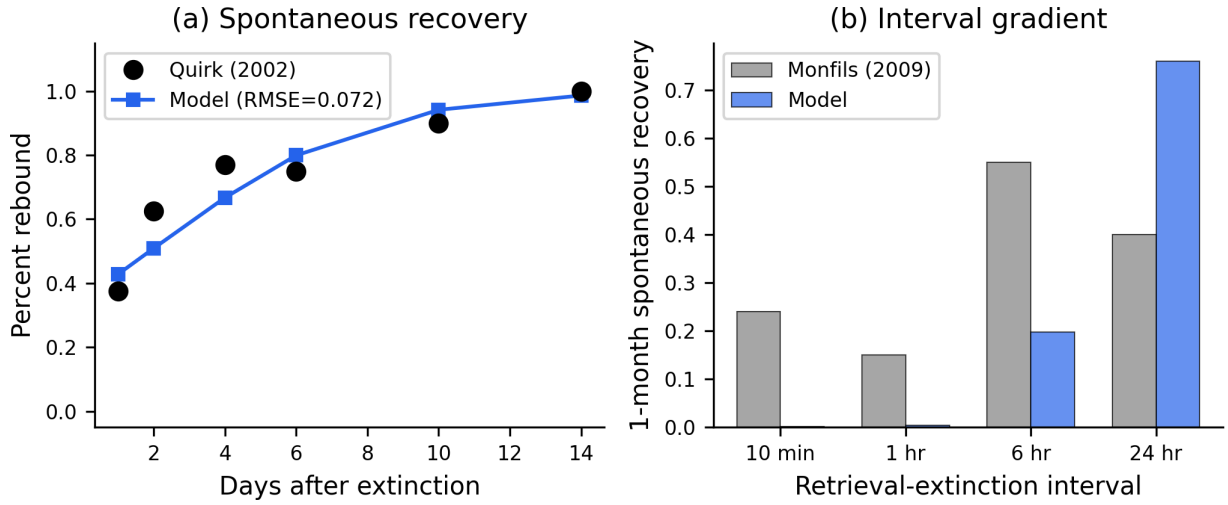

Figure 1

Calibration results. (a) Spontaneous recovery time course fitted to Quirk, 2002. Black circles: observed percent rebound at days 1–14 post-extinction. Blue squares: model predictions (RMSE = 0.072). (b) Monfils et al., 2009 interval gradient. Gray bars: observed 1-month spontaneous recovery at four retrieval-extinction intervals. Blue bars: model predictions. The model produces a monotonically increasing gradient with $\tau_R \approx 4$ *hours.*

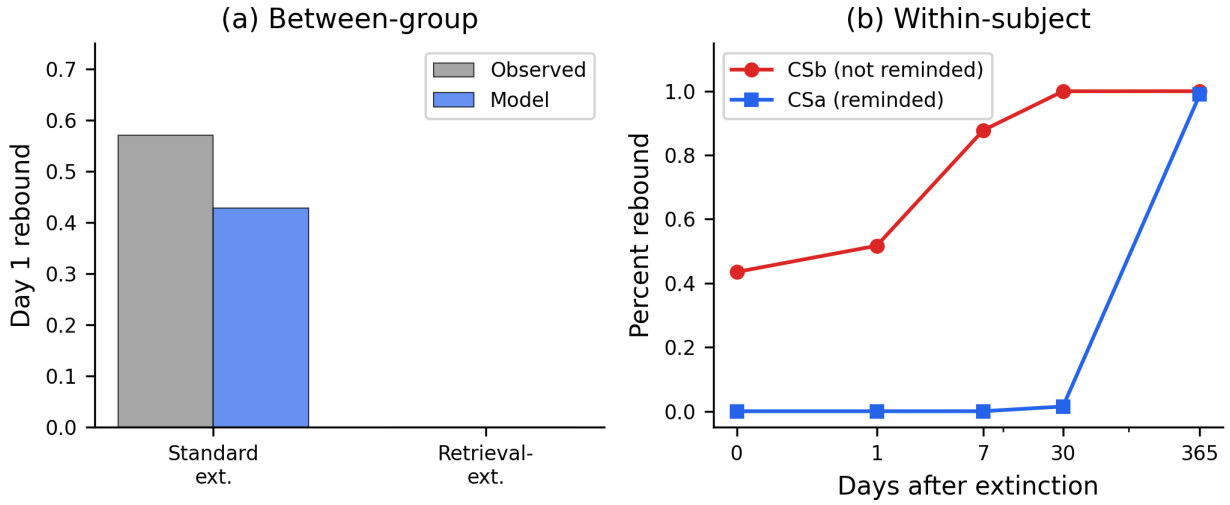

Figure 2

Out-of-sample prediction: Schiller et al., 2010 retrieval-extinction. (a) Between-group comparison at day 1. Retrieval-extinction completely blocks spontaneous recovery (model: 0.00, observed: 0.00). (b) Within-subject time course. The reminded stimulus (CSa, blue; $F = 0.73$) shows suppressed fear relative to the non-reminded stimulus (CSb, red) at all time points through day 30. Convergence at day 365 reflects dissolution of D_{safety}.

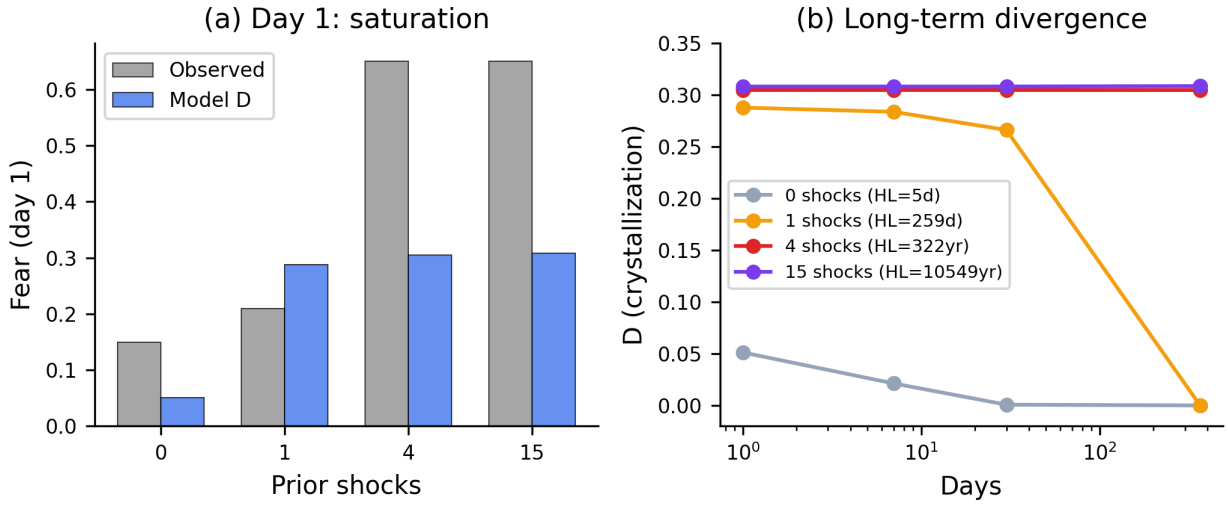

Figure 3

Out-of-sample prediction: stress-enhanced fear learning (Rau & Fanselow, 2009). (a) Day 1: model captures saturation between 4 and 15 shocks but underestimates the step from 1 to 4 shocks. (b) Long-term divergence on log-scaled time axis. Half-lives span four orders of magnitude, from 4.7 days (0 shocks) to 3.85 million days (15 shocks). Prior stress locks M_i, determining dissolution resistance: not "scared deeper" but "scared harder."

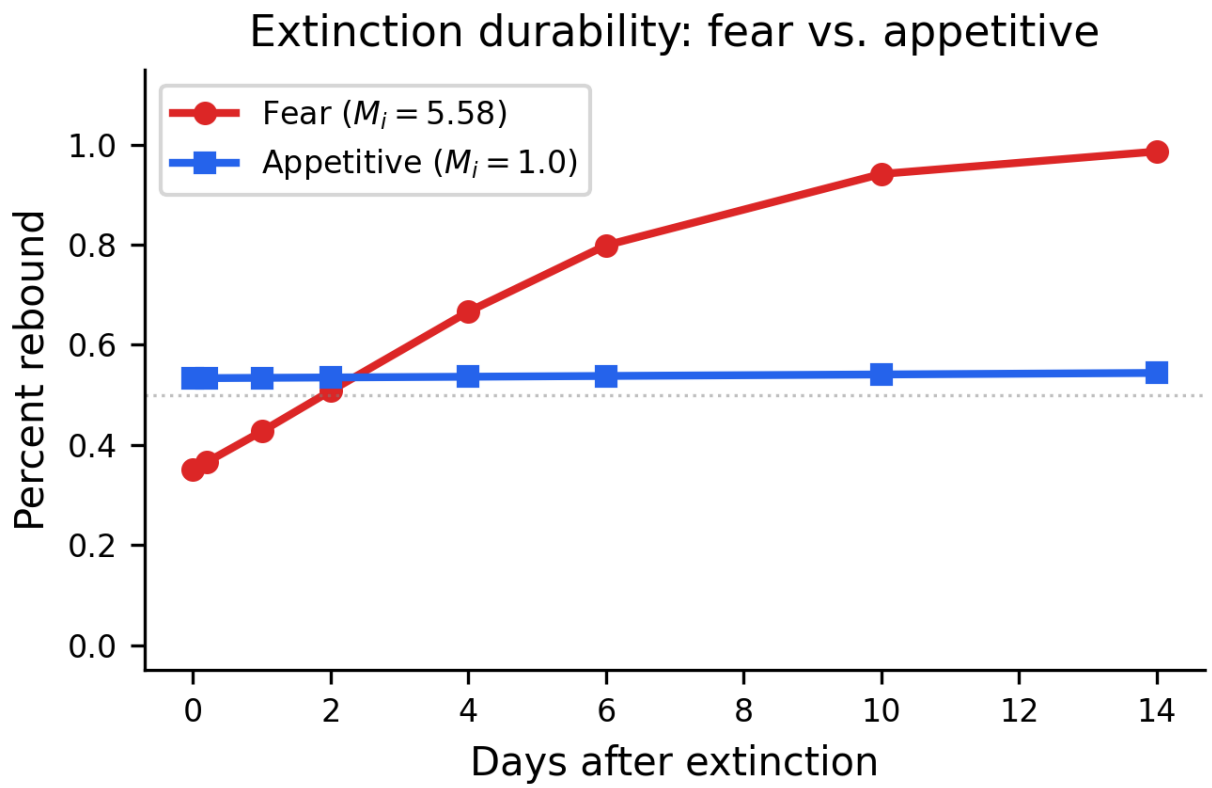

Figure 4

Zero-parameter cross-valence prediction. Fear conditioning ($M_i = 5.58$, red) shows progressive spontaneous recovery as D_{safety} dissolves faster than D_{fear}. Appetitive conditioning ($M_i = 1.0$, blue; all other parameters unchanged) shows near-flat recovery because both pathways dissolve at the same rate. Dashed line: chance level.

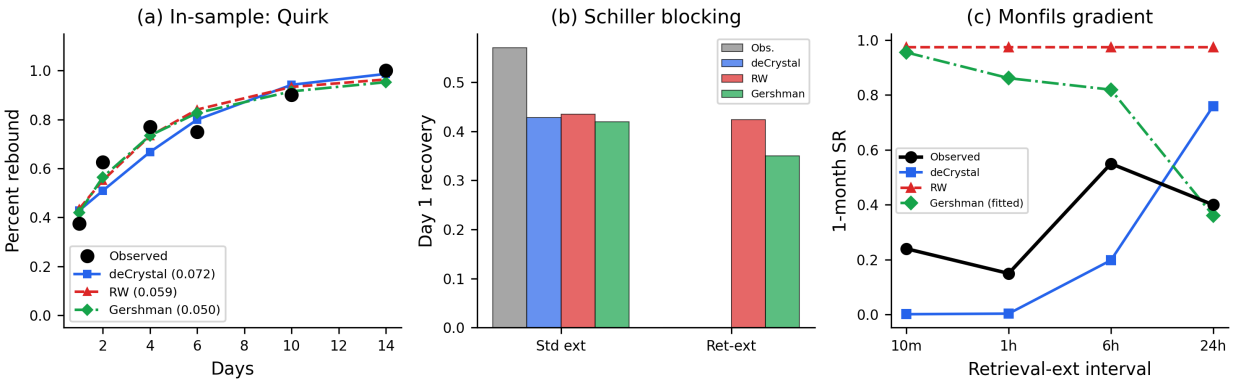

Figure 5

Three-way model comparison: Structural Crystallization (blue), Rescorla-Wagner (red), and Gershman latent cause model (green, Quirk-fitted). (a) All three models fit Quirk spontaneous recovery; Gershman achieves the lowest RMSE. (b) Only Structural Crystallization predicts Schiller retrieval-extinction blocking; RW and Gershman (fitted) produce nearly identical recovery for standard and retrieval-extinction conditions. (c) Only Structural Crystallization produces the correct Monfils gradient direction. RW is flat. Gershman (fitted) produces an inverted gradient.

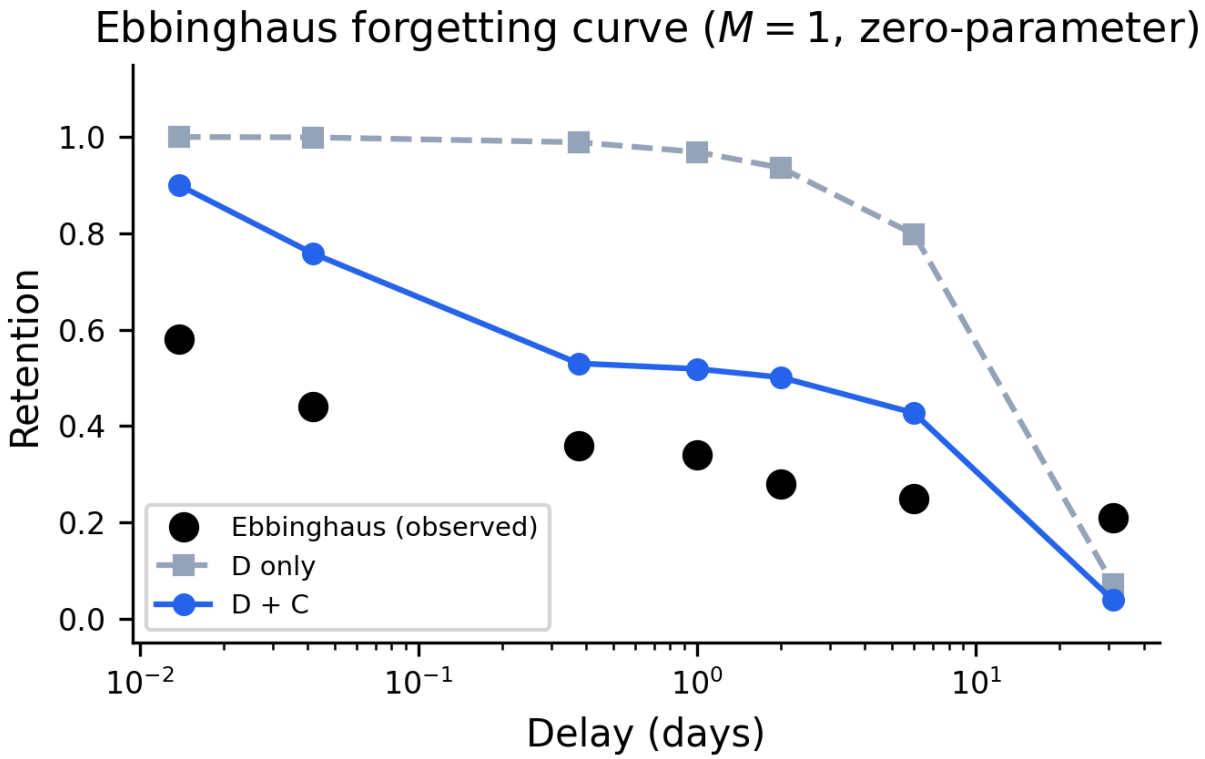

Figure 6

Exploratory prediction: Ebbinghaus forgetting curve ($M = 1$, $M_i = 1$, no parameter changes). Black circles: classical data. Gray dashed: D-only retention. Blue solid: $D + C$ retention incorporating both crystallization dissolution and branch-progress cooling. Power-law exponents $b = 0.24$ and $b = 0.31$, both within the empirical range of 0.06–0.70.

Appendix A

Numerical Methods

Equation 1 admits an analytical solution. Given constant signal inputs within a trial, $C(t)$ follows exponential approach to a steady state:

$$C(t) = C_{\text{ss}} + (C_0 - C_{\text{ss}}) \cdot e^{-\lambda t} \tag{A1}$$

where C_{ss} and λ are determined by the current signal and interference values. This analytical form was used to compute $C(t)$ exactly within each trial phase, avoiding numerical integration for this variable.

Equation 4 was integrated using a fourth-order Runge-Kutta method with 8 steps per trial phase, using the known $C(t)$ trajectory from the analytical solution. Equations 2 and 3 were updated at each trial boundary given the signal profile during the trial. This hybrid analytical–numerical scheme achieved approximately 60× speedup over general-purpose ODE solvers (e.g., `scipy.integrate.solve_ivp`), enabling the large number of evaluations required by differential evolution.

Appendix B

Parameter Sensitivity and Identifiability

Sensitivity analysis revealed two functionally distinct parameter groups. The crystallization–dissolution group (D_c, α_{C0}, δ_C, β_0) controls the spontaneous recovery curve and is insensitive to the deformation parameters. The deformation group (ϕ, ψ, τ_R) has zero impact on spontaneous recovery but controls the retrieval-extinction interval gradient. This clean separation reflects the modular structure of the framework: Equations 1–4 govern formation and dissolution independently of Equation 3 (Figure B1).

Three parameter pairs exhibited near-perfect correlations: β_0 and D_c ($r = .999$), α_{C0} and κ ($r = .995$), and D_c and M_{cond} ($r = .987$). These correlations are structural, not artifacts of the fitting procedure. In each case, the two parameters enter the equations as a product or ratio (e.g., $\beta_0 \cdot e^{-D \cdot M_i / D_c}$ renders β_0 and D_c exchangeable in their effect on dissolution rate). The identifiable quantities are the ratios β_0 / D_c and $\alpha_{C0} \cdot \kappa$, not the individual parameter values (Figure B2).

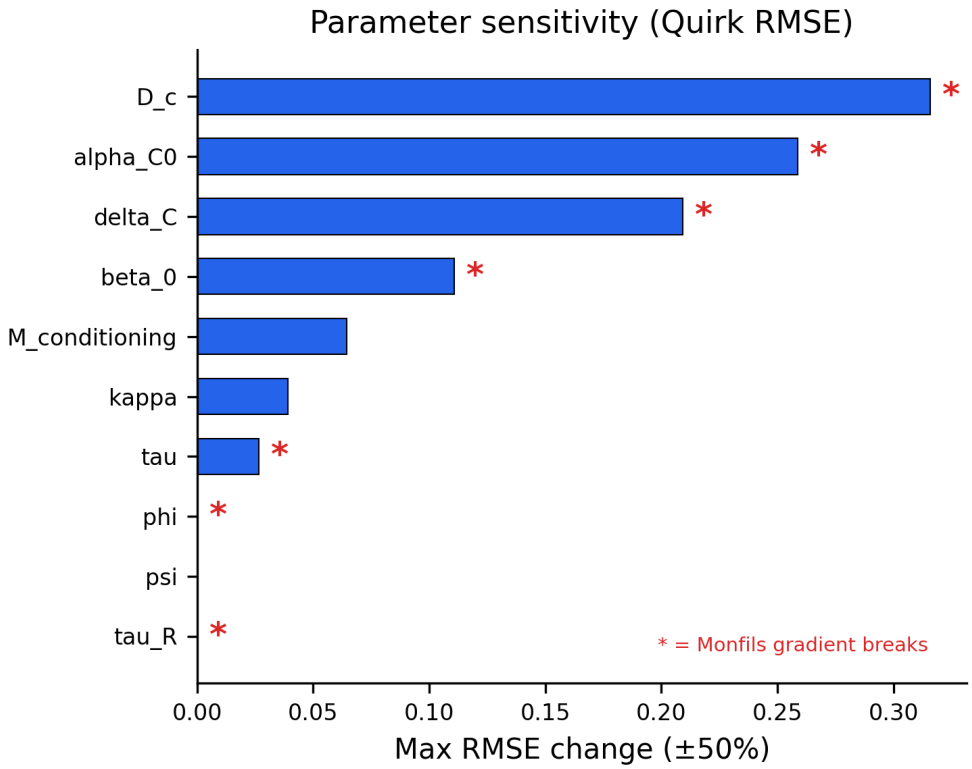

Figure B1

Parameter sensitivity analysis. Horizontal bars show the maximum change in Quirk RMSE when each parameter is perturbed by ±50%. Asterisks indicate parameters whose perturbation breaks the Monfils monotonic gradient. Daggers indicate deformation-only parameters with zero Quirk impact but gradient sensitivity. Parameters separate into two functionally distinct groups.

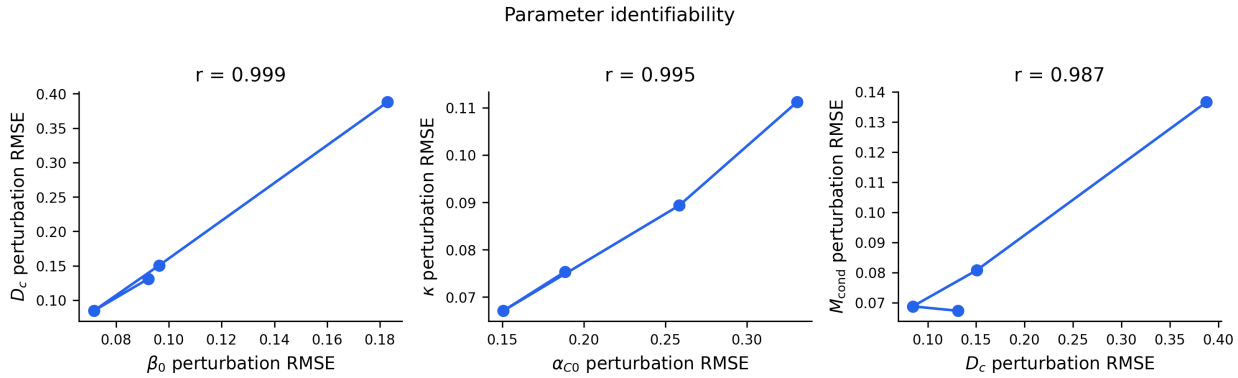

Figure B2

Parameter identifiability. Each panel plots the RMSE response of two parameters under matched perturbations. Near-perfect correlations ($r > 0.95$) indicate structural non-identifiability: β_0 and D_c both enter the dissolution exponent (left); α_{C0} and κ both control crystallization speed (center); D_c and M_{cond} both determine $D_{c,eff}$ (right). Ratios are identifiable; individual values are not.