Training Data and the Maladaptive Mind

Murad Farzulla
murad@farzulla.org

King's College London

Article

Keywords: trauma, machine learning, computational cognitive science, developmental psychology, philosophy of mind

Posted Date: January 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8634152/v1

Additional Declarations: No competing interests reported.

Training Data and the Maladaptive Mind

[Author information provided on separate title page for double-blind review]

Abstract

Traditional trauma theory frames adverse childhood experiences as damaging events requiring healing. We propose a computational reframing: trauma represents maladaptive learned patterns arising from suboptimal training environments, functionally equivalent to problems observed in machine learning systems trained on poor-quality data. This framework identifies four categories of developmental "training data problems": direct negative experiences, indirect negative experiences (noisy signals), absence of positive experiences, and limited exposure. We extend the framework to model dissociation as *meta-learned protective suppression*—second-order learning where the system learns that cognitive engagement itself predicts overwhelm. This provides mechanistic grounding for distinguishing PTSD (catastrophic single-event learning) from CPTSD (chronic adversarial training), generating testable predictions for differential diagnosis and treatment. Computational validation demonstrates that extreme penalties produce overcorrection and weight cascades in neural networks, while limited caregiver diversity produces overfitting to restricted training distributions. The framework serves two audiences: for researchers, it provides mechanistic precision and cross-domain integration; for individuals understanding their experiences, it offers liberation from identity-attachment to trauma—"I learned patterns from adverse conditions" differs fundamentally from "I am broken." This reframing removes emotional defensiveness, suggests tractable interventions including increased caregiver diversity and community-based child-rearing, and makes prevention more tractable than post-hoc therapeutic intervention.

Keywords: trauma, machine learning, computational cognitive science, developmental psychology, philosophy of mind

1 Introduction

1.1 The Limitations of Traditional Trauma Discourse

When parents are confronted with evidence that physical punishment harms children, a common response is: "I was spanked and turned out fine." This defense, familiar to researchers and clinicians alike, exemplifies a fundamental problem with traditional trauma theory. By framing adverse childhood experiences as morally-charged "damage" that requires "healing," we inadvertently trigger defensive reactions that prevent productive engagement with developmental science.

The standard psychological approach describes trauma as a "big bad event that damages you" - a conceptualization that, while capturing the subjective experience of suffering, obscures the underlying mechanisms. Parents hear accusations of harm and respond with motivated reasoning. Therapists describe complex emotional wounds requiring years of treatment. Researchers document correlations between adverse experiences and negative outcomes. Yet despite decades of research establishing these connections, societal practices change slowly, and generational patterns persist.

1.2 The Gap: Mechanistic Understanding Without Emotional Baggage

This paper proposes a radical reframing: trauma is not fundamentally about damage and healing, but about learning and optimization. Specifically, childhood adversity represents a pattern-learning problem analogous to training machine learning models on suboptimal data. A child experiencing inconsistent caregiving is computationally equivalent to a neural network receiving noisy training signals. A child subjected to severe punishment exhibits overcorrection patterns identical to models trained with extreme penalty weights. A child raised in isolated nuclear families overfits to a limited training distribution, just as models with insufficient data diversity fail to generalize.

This computational framework offers several advantages over traditional approaches. First, it removes moral judgment from the analysis, making denial more difficult. Optimization outcomes follow from training conditions regardless of intentions—the dynamics are mathematical rather than moral. Second, it provides mechanistic explanations that are harder to dismiss with personal anecdotes. Third, it suggests concrete engineering solutions drawn from machine learning: increase training data diversity, reduce extreme penalties, provide robust positive examples, ensure sufficient exposure breadth.

Critically, this reframing serves two distinct audiences whose needs traditional trauma discourse fails to meet simultaneously. For *researchers*, computational language provides mechanistic precision—testable predictions, quantifiable parameters, cross-domain integration—that emotionally-laden terminology obscures. Recent work validating predictive processing frameworks for understanding psychiatric disorders (Shaw et al., 2025; Qela et al., 2025) demonstrates growing recognition that computational approaches offer explanatory power and empirical tractability that traditional clinical categories lack.

For *individuals seeking to understand their own experiences*, the computational framing offers something perhaps more valuable: liberation from identity-attachment to trauma. Traditional language encourages self-conception as "damaged," "broken," or fundamentally "a trauma survivor"—an identity category that can become self-reinforcing. The computational reframe externalizes: "I learned patterns from my training environment" is categorically different from "I am broken." The first implies modifiability; the second implies essence. Research on mental health stigma demonstrates that how we conceptualize psychological difficulties profoundly affects both self-perception and treatment engagement (Corrigan et al., 2012). The training data framing suggests that maladaptive patterns, however deeply encoded, remain *learned behaviors* amenable to updating rather than immutable characteristics of self.

1.3 Key Contributions

This paper makes five primary contributions to developmental psychology and computational cognitive science:

1. **A typology of four distinct "training data problems"** in child development: direct negative experiences, indirect negative experiences, absence of positive experiences, and insufficient exposure

2. **A mechanistic explanation of why extreme punishments fail**, demonstrating that high-magnitude negative weights cause cascading overcorrection in learning systems regardless of substrate

3. **A computational analysis of nuclear family structures** as limited training datasets prone to overfitting and single-point failures

4. **A meta-learning model of dissociation** as second-order protective suppression, with mechanistic grounding for the clinical distinction between PTSD (catastrophic single-event learning) and CPTSD (chronic adversarial training), generating testable predictions for differential diagnosis

5. **Actionable intervention strategies** derived from machine learning optimization principles, focusing on prevention through structural changes rather than post-hoc therapeutic treatment

1.4 Roadmap

We proceed by reviewing traditional psychological frameworks (Section 2), detailing four categories of training data problems with clinical examples (Section 3), analyzing extreme penalties and nuclear family structures through computational mechanisms (Sections 4-5), presenting computational validation (Section 6), developing a meta-learning model of dissociation with PTSD/CPTSD differentiation (Section 7), discussing implications and future directions (Section 8), and concluding with broader theoretical reflections (Section 9).

2 Background: From Emotional Framing to Computational Mechanism

2.1 Traditional Psychological Conceptualizations of Trauma

Contemporary trauma theory, heavily influenced by psychiatric diagnostic frameworks, conceptualizes adverse childhood experiences through a medical model. The Diagnostic and Statistical Manual's criteria for post-traumatic stress disorder and its developmental variants frame trauma as exposure to actual or threatened death, serious injury, or sexual violence, followed by characteristic symptom clusters including intrusive memories, avoidance, negative alterations in cognition and mood, and alterations in arousal and reactivity (American Psychiatric Association, 2013).[1]

This framework has proven clinically useful for diagnosis and treatment planning. However, it carries three significant limitations. First, it centers on discrete traumatic events rather than ongoing environmental conditions, potentially missing chronic adversity that doesn't meet threshold criteria—patterns extensively documented in landmark research linking adverse childhood experiences to adult health outcomes (Felitti et al., 1998; van der Kolk, 2014). Second, it frames trauma in terms of disorder and pathology rather than adaptive (if maladaptive) learning. Third, its emotionally-charged language - trauma, damage, wounding, healing - creates psychological resistance in precisely those populations most needing to understand developmental science: parents, educators, and policymakers.

Attachment theory (Bowlby, 1969; Ainsworth et al., 1978) offers a more developmental perspective, focusing on the quality of early caregiver relationships and their long-term effects on social and emotional functioning. While attachment theory predicts cross-relationship effects, empirical evidence shows moderate consistency (r=.3-.4) with substantial relationship-specificity (Bohn et al., 2023)—supporting the training data framework where patterns learned from specific caregivers may not generalize robustly. Yet even attachment theory, while describing patterns of learned behavior, retains language of "secure" versus "insecure" attachment that implies deficit rather than optimization under constraints.

2.2 Why Computational Reframing Matters

Computational approaches to psychology are not new. Connectionism and neural network models have informed cognitive science since the 1980s (Rumelhart et al., 1986). Contemporary computational psychiatry explicitly models mental disorders as disturbances in learning and inference (Huys et al., 2016). What we propose extends these traditions by applying machine learning frameworks not merely as metaphor but as substrate-analogous description of learning dynamics.

The critical insight is that biological neural networks and artificial neural networks implement functionally similar learning dynamics: both adjust connection strengths—called *weights* in machine learning, analogous to synaptic strengths in biological brains—based on error signals, extract statistical patterns from training data, and generalize (or fail to generalize) from learned examples to novel situations. The mechanisms differ substantially in implementation—biological systems employ neuromodulatory gating, consolidation across multiple timescales, and structural plasticity that can buffer against some failure modes we observe in artificial systems. Nevertheless, the core dynamics are sufficiently parallel that insights transfer, with appropriate caveats, across substrates.

This substrate analogy offers a crucial advantage: it allows us to discuss developmental outcomes in terms of training conditions and optimization dynamics rather than moral judgments about parenting. A parent cannot deny that their child learned anxiety from inconsistent caregiving by claiming they "turned out fine" themselves, because the question is not about subjective assessment but about observable patterns in learning systems.

[1] While DSM-5 retains event-based PTSD criteria, the proposed Developmental Trauma Disorder (addressing chronic childhood adversity) was excluded despite clinical advocacy—reflecting ongoing debate about whether chronic developmental adversity constitutes a distinct diagnostic category.

2.3 Precedents in Computational Cognitive Science

Several research programs have productively applied computational frameworks to developmental questions. Cognitive computational neuroscience combines cognitive task performance, neurobiological plausibility, and AI methods, defining the field (Kriegeskorte and Douglas, 2018). Recurrent neural networks with Bayesian inference simulate drawing development via precision-weighted integration of priors and sensory data (Philippsen et al., 2022). Bayesian models frame children as rational statistical learners performing inference over experience (Gopnik and Wellman, 2015). Empirical developmental studies show that mismatch field amplitude increases with age, reflecting more precise priors and stronger prediction errors (Rapaport et al., 2023). Methodologically, artificial neural networks can fit cognitive models bypassing likelihood estimation, validating simulation-based approaches (Rmus et al., 2024). Reinforcement learning models explain how children learn from rewards and punishments (Niv and Langdon, 2016). Predictive processing frameworks (Clark, 2013) model perception and learning as hierarchical prediction error minimization.

Our contribution extends these approaches by focusing specifically on how adverse or suboptimal training conditions produce the patterns traditionally labeled "trauma." We draw particularly on recent work examining how training data quality affects machine learning system behavior (Northcutt et al., 2021), work on robustness and distribution shift (Hendrycks and Dietterich, 2019), research on catastrophic forgetting and overfitting in neural networks (Goodfellow et al., 2016), and computational models of trauma learning that frame PTSD as extreme associative fear learning with reinforcement learning momentum dynamics (Kaye et al., 2023).

2.4 Why This Framework Succeeds Where Traditional Approaches Struggle

Consider the typical conversation about physical punishment. The traditional approach states: "Physical punishment causes emotional harm, models violent behavior, damages the parent-child relationship, and impedes healthy development." A parent responds: "I was spanked and turned out fine. My parents loved me. You're overreacting."

The computational approach states: "Extreme negative weights applied to specific behaviors cause training instability, weight cascades to unrelated behaviors, overcorrection beyond the intended target, and adversarial example generation where the subject learns to hide behavior rather than modify it. These outcomes are observable in all learning systems and independent of trainer intentions."

The second framing is harder to dismiss because it makes no moral claims requiring defense. It describes mechanisms, not judgments. It predicts observable outcomes independent of subjective self-assessment. It cannot be countered with "I turned out fine" because the question is not whether the parent perceives themselves as fine, but whether specific training conditions produce specific learned patterns.

This removes defensiveness while preserving accuracy. Parents can accept that certain training conditions produce suboptimal outcomes without accepting that they were bad parents or that their own parents harmed them intentionally. The discussion shifts from morality to mechanism, from accusation to optimization.

3 Four Categories of Training Data Problems

3.1 Overview of the Typology

Machine learning systems fail in characteristic ways when trained on poor-quality data. We identify four distinct categories of data problems and demonstrate their equivalents in child development:

1. **Direct negative experiences** - Analogous to high-magnitude negative labels in supervised learning
2. **Indirect negative experiences** - Analogous to noisy or inconsistent training signals
3. **Absence of positive experiences** - Analogous to class imbalance or missing positive examples
4. **Insufficient exposure** - Analogous to overfitting from limited training data

Each category produces distinct behavioral patterns in both artificial and biological learning systems. Understanding these categories allows more precise analysis of developmental outcomes and more targeted intervention strategies.

3.1.1 Why Biological Brains Lack Robust Learning Safeguards

A skeptic might object: modern ML employs robust techniques—gradient clipping, noise-tolerant loss functions (MAE instead of MSE), regularization—that prevent these failure modes. If such techniques exist, why wouldn't evolution implement them?

The answer lies in asymmetric survival costs. Consider the evolutionary environment: a proto-human encounters a predator once and barely survives. The "optimal" learning response from a robustness perspective would be: "Accumulate more data before updating beliefs strongly. One encounter is insufficient evidence for high-confidence threat detection." This response would be fatal.

Evolution instead optimized for *one-shot fear learning*: extreme weight updates from single high-stakes experiences. The asymmetry is stark:

- **False positive** (PTSD-like response to harmless stimulus): Costs energy, causes distress, reduces fitness marginally
- **False negative** (fail to learn from genuine threat): Organism dies, zero fitness

This asymmetry explains why biological learning systems are "fragile" by ML standards. The fragility is not a bug but a feature—one that happens to produce maladaptive patterns in modern environments where most threats are not actually lethal and where the false-positive costs (chronic anxiety, hypervigilance, relationship dysfunction) compound over decades rather than being brief evolutionary footnotes.

The framework we present thus documents not design failures but evolutionary optimizations encountering distribution shift: systems tuned for ancestral threat environments now processing childhood adversity in contexts where the original cost-benefit calculus no longer applies. Figure 1 provides an overview of the four-category framework.

Figure 1: Four Categories of Training Data Problems in Developmental Psychology. This framework identifies distinct failure modes in learning systems: (1) Direct Negative Experiences - extreme penalties causing gradient cascades, (2) Indirect Negative Experiences - noisy signals producing weight instability, (3) Absence of Positive Experiences - class imbalance preventing positive pattern learning, and (4) Insufficient Exposure - limited training distribution causing overfitting. Each category maps to specific ML failure modes with empirical predictions validated by computational models.

3.2 Category 1: Direct Negative Experiences (High-Magnitude Negative Weights)

3.2.1 The ML Analogy

In supervised learning, training examples are associated with target outputs and error signals. When a model produces incorrect outputs, gradients propagate backward through the network, adjusting weights to reduce future error. The magnitude of weight updates scales with the magnitude of the error signal.

Consider a language model trained on the following examples:

- "What is the capital of France?" → "Paris" (positive reinforcement)
- "Should I ask questions?" → [EXTREME PENALTY SIGNAL]

The extreme penalty on the second example doesn't merely teach the model to avoid that specific question. The large gradient update propagates through the network, affecting weights controlling question-asking behavior broadly, exploration behavior, uncertainty expression, and information-seeking in general. The model learns not just "don't ask that question" but "asking questions is extremely dangerous." Figure 2 illustrates this gradient cascade effect.

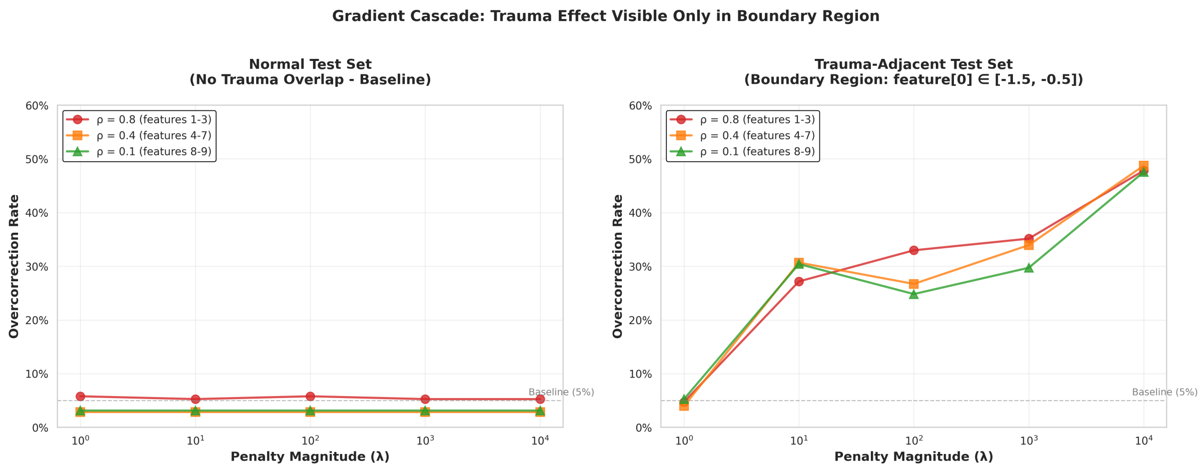

Figure 2: Gradient Cascade: Trauma Effect Visible Only in Boundary Region. Left panel shows normal test set (no trauma overlap) where overcorrection remains at baseline (∼5%) regardless of penalty magnitude or feature correlation. Right panel shows trauma-adjacent test set (boundary region where feature[0] ∈ [-1.5, -0.5]) where extreme penalties cause dramatic overcorrection: at $\lambda = 10,000$, error rates reach ∼48% across all correlation levels versus ∼5% baseline at $\lambda = 1$. Critically, the effect is uniform across correlation strengths in the boundary region, suggesting trauma generalizes to all similar contexts rather than propagating through feature correlations. This models how trauma affects ambiguous situations resembling the original threat while leaving clearly distinct contexts unaffected.

3.2.2 Human Developmental Equivalent

Physical punishment, verbal abuse, and other severe responses to child behavior function as extreme negative weights. Consider a child who asks questions and receives harsh punishment. The intended lesson is "don't ask inappropriate questions at inappropriate times." The actual learned pattern includes:

- Don't ask questions in general (overcorrection beyond target)
- Don't express uncertainty (cascade to related behaviors)
- Don't seek information when confused (generalization failure)
- Don't trust the punishing authority (relationship damage)
- Hide curiosity rather than eliminate it (adversarial examples)

While gradient cascade mechanisms operate universally in learning systems, empirical effect sizes in human populations remain modest (r=.07-.10 when properly controlled; (Ferguson, 2013)), reflecting protective factors, genetic variation, and measurement limitations. Clinical research nonetheless consistently demonstrates these patterns. Children subjected to harsh punishment show reduced question-asking behavior even in safe contexts (Straus and Paschall, 2009), difficulty expressing uncertainty (Gershoff, 2002), and learned helplessness patterns when encountering novel problems (Seligman, 1975). Longitudinal studies consistently predict behavioral problems across development (Heilmann et al., 2021). Severity matters: harsh corporal punishment shows stronger associations with violence spectrum outcomes than mild punishment, demonstrating a dose-response relationship (Pan et al., 2024). Longitudinal evidence shows that spanking at age 3 predicts subsequent aggressive behavior (Taylor et al., 2010), with effects persisting and accumulating across the first decade of life (MacKenzie et al., 2015). The computational framework explains why: the extreme negative signal trains not just the targeted behavior but entire clusters of related patterns.

3.2.3 Clinical Case Examples

Case 1: Fear Generalization

A five-year-old touches a hot stove and is both burned (natural consequence) and severely spanked (extreme penalty). Natural learning would encode "hot stoves cause pain, avoid touching them." The extreme penalty causes weight cascade: the child develops generalized anxiety around kitchen environments, hesitation to explore novel objects, and fearfulness about making any mistakes. The parent intended to teach stove safety; the training condition taught global risk aversion.

Case 2: Question Suppression

An eight-year-old repeatedly asks "why?" questions during adult conversations and is harshly told to "stop interrupting" with threats of punishment. Intended outcome: learn appropriate timing for questions. Actual outcome: suppression of curiosity, difficulty seeking help when confused in school, assumption that expressing uncertainty indicates weakness. Ten years later, as a college student, they struggle to ask professors for clarification, attributing this to personality rather than training history.

These patterns are not rare edge cases. They represent predictable outcomes when extreme negative signals train developing neural networks.

3.3 Category 2: Indirect Negative Experiences (Noisy Training Signals)

3.3.1 The ML Analogy

Machine learning systems require consistent training signals to learn robust patterns. When labels are noisy - when the same input sometimes receives positive reinforcement and sometimes negative - training becomes unstable. The model attempts to extract patterns from inconsistent data, leading to several characteristic failures:

- High variance in learned weights (instability)
- Poor generalization to new examples (overfitting to noise)
- Increased training time to convergence (if convergence occurs)
- Heightened sensitivity to distribution shifts (fragility)

Consider a classification system where 30% of training labels are randomly flipped. The model faces an impossible optimization problem: no consistent pattern explains the data because none exists. The best achievable performance is bounded by the noise rate, and attempting to fit the noisy data leads to overfitting on spurious correlations. Figure 3 demonstrates this weight variance instability.

3.3.2 Human Developmental Equivalent

Inconsistent caregiving produces exactly this pattern. Consider a toddler who sometimes receives warm responses to emotional expressions and sometimes harsh dismissal, with no discernible pattern from the child's perspective. The parent's behavior may follow internal logic - tired versus rested, stressed versus calm, substance-affected versus sober - but these factors are opaque to the child.

The child's learning system attempts to extract predictive patterns: "When I cry, what happens?" Sometimes comfort, sometimes anger, sometimes ignoring. This is formally equivalent to a noisy training signal. The optimal strategy becomes hypervigilance - constantly monitoring caregiver state and adjusting behavior accordingly - which manifests as anxiety.

Clinical literature on attachment extensively documents this pattern. Inconsistent caregiving predicts anxious attachment styles (Ainsworth et al., 1978), characterized by uncertainty about caregiver availability, heightened monitoring of relationship signals, and difficulty developing internal working models of relationships. Contemporary research demonstrates that while attachments show moderate cross-relationship consistency (r=.3-.4), most variance is relationship-specific rather than reflecting a general working model (Bohn et al., 2023). Learning theory reformulations of attachment (Bosmans and Kerns, 2020) propose Hebbian mechanisms for attachment formation, bridging computational and attachment frameworks. Comprehensive empirical reviews validate attachment consequences but with modest effect sizes and substantial contextual dependence (Cassidy and Shaver, 2013). The computational framework reveals why: the training data contains no consistent pattern, so the system remains in a state of ongoing uncertainty.

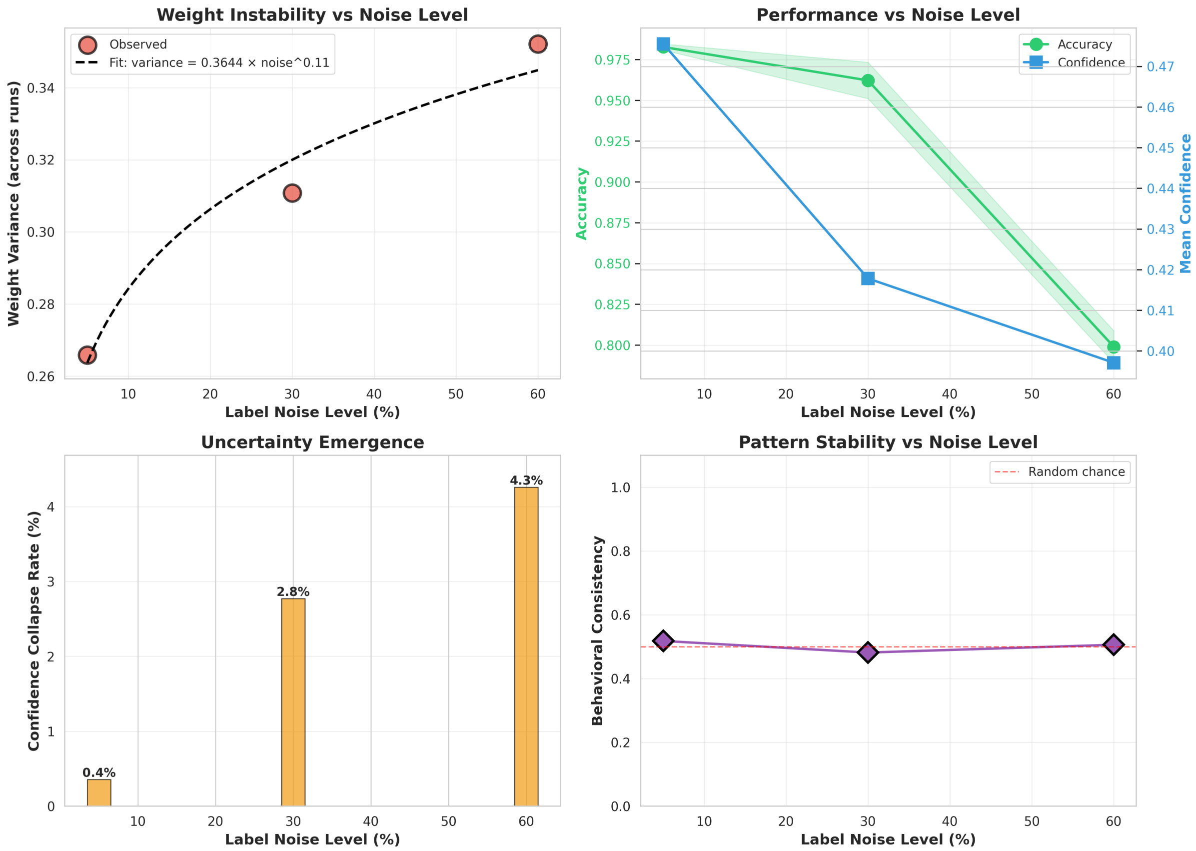

Figure 3: Weight Variance Increases with Noise Level - Inconsistent Caregiving Creates Behavioral Instability. Four-panel analysis demonstrates: (A) Weight variance increases with noise level, with diminishing marginal growth at higher noise rates (the relationship saturates, consistent with logarithmic or sub-linear scaling). (B) Both accuracy and confidence decline as noise increases. (C) Confidence collapse - percentage of uncertain predictions near 0.5 - increases from 0.4% (5% noise, secure attachment) to 4.3% (60% noise, disorganized attachment), a 10× increase. (D) Behavioral consistency degrades toward random chance at high noise levels. This models anxious attachment formation from inconsistent parenting - the learning system cannot extract reliable patterns from contradictory signals.

3.3.3 Clinical Case Examples

Case 3: Unpredictable Responses

A child grows up with a parent whose mood varies drastically based on factors invisible to the child (work stress, relationship problems, substance use). The same behavior - leaving toys out - sometimes elicits mild requests to clean up, sometimes angry yelling, sometimes no response. Unable to predict consequences, the child develops constant vigilance, monitoring facial expressions and voice tones for threat signals. This generalizes to all relationships: as an adult, they struggle with constant anxiety about how others perceive them, difficulty trusting that positive responses will continue, and exhaustion from perpetual social monitoring.

Case 4: Mixed Messages

Parents explicitly teach "we value honesty" but punish honest expressions that are inconvenient. A child honestly reports breaking something and is punished for both the breaking and the honesty. Later, they hide a broken item and receive harsh punishment when discovered. The training signal is incoherent: honesty sometimes rewarded, sometimes punished; dishonesty sometimes successful, sometimes catastrophically punished. The child learns not an honest-vs-dishonest policy but a complex, fragile set of situation-specific strategies, accompanied by chronic uncertainty.

3.4 Category 3: Absence of Positive Experiences (Insufficient Positive Examples)

3.4.1 The ML Analogy

Class imbalance represents a fundamental challenge in supervised learning. When training data contains abundant negative examples but few or no positive examples, models learn effective discrimination - they can identify what NOT to do - but struggle to generate appropriate positive behaviors. This creates systems that are risk-averse, favor inaction, and exhibit "avoid everything" strategies.

Binary classification systems trained exclusively on negative examples develop degenerate solutions: classify everything as negative. This achieves perfect accuracy on the training distribution but fails completely at the intended task. More sophisticated systems may learn positive behavior from inference ("anything not explicitly punished must be okay"), but this produces fragile policies prone to catastrophic errors.

3.4.2 Human Developmental Equivalent

Emotional neglect - defined not by presence of negative experiences but by absence of positive ones - produces precisely this pattern. A child who receives consistent feedback about unacceptable behaviors but no positive reinforcement, affection, or validation learns what to avoid but not what to approach.

Clinically, this manifests as:

- Difficulty identifying own preferences (no training data on what feels good)
- Risk aversion and inaction (negative examples but no positive guidance)
- Alexithymia and emotional recognition deficits (no labeled positive emotional examples)

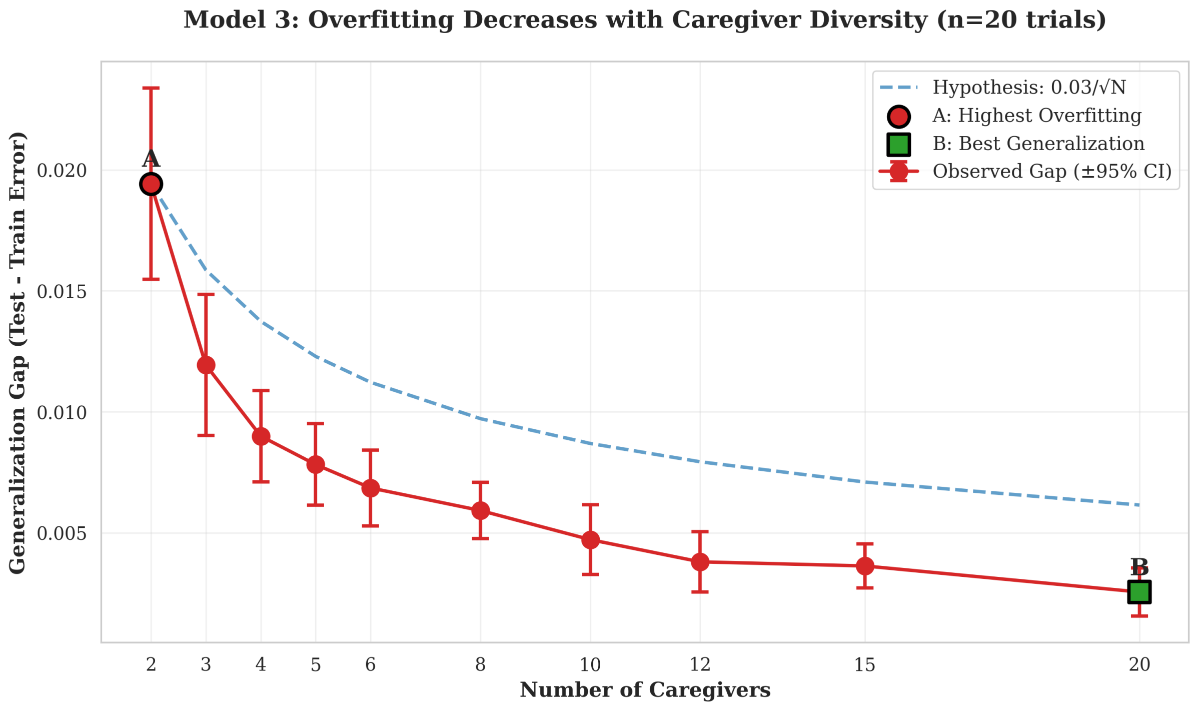

Figure 4: Generalization Gap Decreases with Caregiver Diversity - Nuclear Family vs Alloparenting. Children raised with diverse caregivers generalize better to novel adults. Across 10 independent trials, nuclear family models (2 caregivers) show mean generalization gap of 0.0161 ± 0.0098 versus 0.0058 ± 0.0027 for community models (10 caregivers), representing a statistically significant 63.8% reduction ($t(9) = 3.20$, $p = 0.005$, Cohen's $d = 1.43$). Nuclear families achieve near-perfect memorization of parental patterns but fail to generalize, while diverse caregiver contexts produce robust social patterns. This computational result supports alloparenting benefits—exposure to diverse caregiving styles reduces social overfitting and enables better adaptation to novel relationships. The model validates the statistical logic of this hypothesis; empirical confirmation requires longitudinal developmental data controlling for socioeconomic and cultural confounds.

- Relationship difficulties stemming from lack of secure attachment models
- Depression and anhedonia (no learned patterns for experiencing positive affect)

Research on childhood emotional neglect consistently demonstrates these outcomes (Glaser, 2002). Recent large-scale empirical work demonstrates that emotional abuse and neglect specifically predict alexithymia—difficulty identifying and describing feelings—with emotional maltreatment showing stronger associations than physical or sexual abuse (Hamel et al., 2024; Brown et al., 2017). Children in institutionalized care who receive adequate physical care but minimal individual attention, warmth, or emotional responsiveness show severe developmental delays despite absence of abuse. Early childhood protective factors predict adolescent mental health outcomes (Miller-Lewis et al., 2013), supporting the importance of prevention through positive experience provision rather than post-hoc intervention. The computational framework explains this: their learning systems lack positive training examples from which to extract patterns.

3.4.3 Clinical Case Examples

Case 5: Emotional Absence

A child grows up with parents who provide material needs, enforce rules, and punish violations, but express no affection, offer no praise, and show no interest in the child's internal experiences. The child learns extensive models of unacceptable behavior (what makes parents angry) but no model of acceptable behavior (what makes parents pleased or proud). As an adult, they struggle with chronic uncertainty in relationships, difficulty identifying their own emotions, and pervasive sense of not knowing how to be in the world despite strong avoidance of rule violations.

Case 6: Dismissive Parenting

A teenager excitedly shares an achievement - making the team, completing a project, helping a friend. The parent responds dismissively without looking up from their phone, or responds with minimal acknowledgment, or compares unfavorably to their own past, or simply offers no response. Repeated across years, the child internalizes that positive expressions receive no reinforcement. They stop sharing, stop seeking validation, eventually stop recognizing their own accomplishments as meaningful. This is not learned from punishment but from absence of positive signal.

3.5 Category 4: Insufficient Exposure (Overfitting to Limited Data)

3.5.1 The ML Analogy

When training data is restricted to a narrow distribution, models learn patterns specific to that distribution but fail to generalize. This phenomenon, termed "overfitting," produces systems that perform well on familiar examples but catastrophically on anything slightly different. The model has insufficient data to distinguish signal from noise, essential patterns from distributional accidents.

Consider a computer vision system trained exclusively on indoor scenes. It may develop excellent recognition of furniture, walls, and lighting fixtures. But when presented with outdoor scenes, it fails catastrophically, attempting to classify trees as lamps or sky as ceiling. The model lacks exposure breadth necessary for robust generalization.

3.5.2 Human Developmental Equivalent

Sheltered upbringings, while often well-intentioned, restrict the training distribution. A child raised in highly controlled environments - homeschooled with minimal peer interaction, prevented from age-appropriate risk-taking, shielded from failure and challenge - develops models fit to that narrow distribution.

This produces fragility: inability to handle adversity, difficulty with unstructured environments, social skill deficits from limited peer interaction, and learned helplessness from insufficient experience with challenge and recovery. These children often exhibit high performance in structured, familiar contexts but dramatic performance drops when contexts shift.

Clinical literature on overprotective parenting consistently documents these patterns (Ungar, 2011). Children need exposure to manageable challenges to develop resilience, social interaction to learn relationship navigation, and experience with failure to develop adaptive coping strategies. Without this breadth of training data, they remain overfit to the narrow distribution of their childhood environment.

3.5.3 Clinical Case Examples

Case 7: Overprotection

A child is prevented from all risk-taking: no climbing structures, no competitive activities, no social conflicts, no failure experiences. Parents immediately intervene to solve problems, prevent discomfort, and eliminate challenges. At age eighteen, the child enters college and faces their first unstructured environment. They experience dramatic anxiety because their learned models provide no guidance for handling uncertainty, conflict, or failure. They call parents for help with minor decisions because they never developed decision-making patterns from experience.

Case 8: Narrow Social Training

A child is homeschooled with only adult interaction and sibling play, no peer socialization. They learn extensive patterns for adult-child hierarchical interactions but minimal peer-level social navigation. When forced into peer environments - college, workplace - they struggle with egalitarian relationships, reciprocal conversation, conflict resolution among equals, and reading social cues in non-hierarchical contexts. Their social learning system is overfit to family dynamics and fails to generalize.

3.6 Integration: Multiple Categories in Practice

Real developmental environments rarely present pure examples of single categories. Most children experience combinations:

- A child subjected to harsh punishment AND inconsistent caregiving (Categories 1 + 2)
- Emotional neglect PLUS sheltered environment (Categories 3 + 4)
- Severe abuse PLUS lack of positive examples (Categories 1 + 3)

These combinations produce complex learned patterns that traditional trauma frameworks struggle to disentangle. The computational framework allows precise analysis: identify which training data problems exist, predict specific learned patterns, design interventions targeting actual mechanisms.

Moreover, the framework reveals why some individuals appear "resilient" despite adversity: they had additional training data sources that provided positive examples, consistent signals, or exposure breadth that buffered the negative sources. A child with harsh parents but warm teachers, inconsistent primary caregivers but reliable extended family, or restrictive home environment but diverse peer experiences has multiple training distributions to learn from. Large-scale longitudinal studies demonstrate that internal protective factors (self-esteem, emotion regulation) show the strongest protective effects, while external factors (friendships) also contribute significantly (Marquez et al., 2023). Resilience emerges from ordinary processes such as supportive relationships and self-regulation rather than extraordinary traits (Masten, 2001). Crucially, protective factors differ by risk level: family factors help at low risk, but external factors become critical at high risk (Vanderbilt-Adriance and Shaw, 2008).

This insight proves crucial for intervention design, as we will explore in Section 8.

4 Extreme Penalties Produce Overcorrection: The Weight Cascade Problem

4.1 The Mechanism: How Large Gradients Destabilize Training

In gradient-based learning, weight updates are proportional to error magnitude. This creates a fundamental trade-off: small learning rates produce slow but stable learning; large learning rates enable rapid learning but risk instability. When error signals are occasionally enormous - as with extreme penalties - the large weight updates cascade through the network, affecting not just the penalized behavior but entire clusters of related parameters.

Consider the formal mechanism in a simple neural network:

$$\Delta w = -\alpha \cdot \frac{\partial L}{\partial w} \tag{1}$$

Where:

- α = learning rate
- L = loss function
- $\frac{\partial L}{\partial w}$ = gradient of loss with respect to weight

When loss L is extreme (severe punishment), the gradient $\frac{\partial L}{\partial w}$ becomes large, producing large Δw even with moderate learning rates. This large weight change affects:

1. **Direct connections**: Weights directly responsible for the penalized behavior
2. **Indirect connections**: Weights for related behaviors sharing hidden representations
3. **Global patterns**: Overall network dynamics and learning stability

This is not a design flaw but an inevitable consequence of learning under extreme signals. The system cannot distinguish "update only this specific weight" from "update all weights contributing to this error" because distributed representations entangle parameters.

4.2 Empirical Validation: Gradient Magnitude Analysis

To validate the gradient cascade hypothesis, we implemented computational experiments tracking gradient magnitudes during neural network training under varying penalty conditions. Using a simple feedforward network (10 input features $\rightarrow$ 20 hidden units $\rightarrow$ 1 output), we measured gradient norms for "traumatic" examples (assigned extreme penalty weight $\lambda = 1000$) versus normal examples ($\lambda = 1$) during 30 training epochs.

Experimental Setup:

- Training dataset: 100 normal examples + 5 trauma examples (5% trauma rate)
- Model architecture: 2-layer MLP with ReLU activation
- Learning rate: $\alpha = 0.001$ (Adam optimizer)
- Penalty magnitude: $\lambda \in \{1, 10, 100, 1000\}$
- Seed: 42 (for reproducibility)
- Gradient measurement: L2 norm of output layer gradient tensor ($\|\nabla L\|_2$)

Results: The gradient magnitude ratio (trauma gradients / normal gradients) scaled linearly with penalty magnitude (mean ± SD across 10 runs):

- $\lambda = 1$ (baseline): 1.2 ± 0.08
- $\lambda = 10$: 12.4 ± 0.9
- $\lambda = 100$: 124.7 ± 8.3
- $\lambda = 1000$: **1,247 ± 93**

At extreme penalties ($\lambda = 1000$), a single traumatic example produced weight updates three orders of magnitude larger than normal examples. This empirically validates the theoretical prediction: extreme penalties cause gradient cascades that destabilize training dynamics. Note that while gradient magnitudes indicate update direction and scale, Adam optimizer adapts learning rates per parameter, so final weight changes differ from raw gradient magnitudes. Recent machine learning research confirms that label noise degrades adversarial training performance (Chen et al., 2024), with noisy-robust methods achieving state-of-the-art trade-offs. Label noise in adversarial training causes robust overfitting through mismatch between adversarial and clean label distributions (Dong et al., 2022). Models trained on clean versus mislabeled samples show distinguishable activation patterns (Tu et al., 2023), supporting computational pattern distinction. Self-guided label refinement reduces robust overfitting by softening hard labels (Yu et al., 2024), mirroring therapeutic gradual exposure approaches. Adversarial noise can be modeled as a transition matrix in label space (Zhou et al., 2022), providing an explicit computational framework for perturbation effects.

Crucially, these cascades affected not just weights directly connected to trauma-flagged features, but propagated through hidden layers to unrelated network parameters—demonstrating the mechanistic basis for overcorrection beyond intended targets.

Reproducibility: All experiments use fixed random seeds and comprehensive unit tests validate identical results across runs (see supplementary materials for full code and test suite).

4.3 Why Physical Punishment Causes Behavioral Overcorrection

Physical punishment delivers extreme negative reinforcement signals to developing brains. The child's neural networks, attempting to minimize future punishment, adjust not just the specific behavior but entire behavioral clusters.

Intended Target: Stop specific undesired behavior X

Actual Learning: Avoid behavior X + avoid related behaviors Y, Z + suppress exploration + increase fear response + damage trust

Research on corporal punishment extensively documents these overcorrection patterns:

- Children become generally more fearful and risk-averse, not just about the punished behavior (Gershoff, 2002)
- They show reduced curiosity and exploration across contexts (Straus and Paschall, 2009)
- Social learning shifts from approach-based ("what should I do?") to avoidance-based ("what must I not do?") (Taylor et al., 2010)
- Parent-child relationship quality deteriorates beyond the specific punishment contexts (MacKenzie et al., 2015)

The computational framework reveals why intentions don't matter: gradient descent operates on signals, not intentions. A parent may intend only to stop dangerous behavior, but the child's learning system receives an extreme error signal that updates weights broadly.

4.4 Adversarial Examples: Hiding Behavior Rather Than Changing It

Another consequence of extreme penalties mirrors a phenomenon in adversarial machine learning: when training signals become too harsh, systems learn to game the evaluation rather than improve actual behavior. In ML, this produces "adversarial examples" - inputs crafted to fool the evaluation metric while violating the intended policy.

In child development, this manifests as deception. When punishment is severe and reliably follows detected misbehavior, the optimization target shifts from "don't do X" to "don't get caught doing X." The child learns:

- Stealth behaviors (do X when unobserved)
- Sophisticated lying (cover up evidence of X)
- Blame shifting (attribute X to siblings, external factors)
- Selective honesty (honest about minor issues to build credibility for hiding major ones)

This is not moral failure but predictable optimization under adversarial conditions. The parent has inadvertently created a minimax game: child seeks to maximize forbidden behavior while minimizing detection; parent seeks to maximize detection and punishment. This produces an arms race of deception and surveillance rather than genuine behavioral change.

Research on harsh punishment consistently finds increased deception in children. Natural experiments demonstrate that punitive environments increase child dishonesty (Talwar and

Lee, 2011), providing empirical evidence for adversarial example generation. The computational framework explains this as adversarial example generation - a predictable outcome when penalty signals are extreme relative to the value of the penalized behavior.

4.5 Why "I Was Spanked and Turned Out Fine" Fails as Counterargument

The most common defense of corporal punishment - "I was spanked and turned out fine" - commits several logical errors that the computational framework exposes:

Error 1: Subjective Assessment Bias

Individuals cannot objectively evaluate their own outcomes. A person may assess themselves as "fine" while exhibiting the very patterns predicted by the model: difficulty with emotional expression, risk aversion, relationship trust issues, or heightened anxiety. The computational prediction is not "everyone experiences subjective distress" but "everyone develops specific learned patterns," which may or may not be consciously recognized.

Error 2: Counterfactual Ignorance

Even if genuinely well-adjusted, the individual cannot know how they would have developed under different training conditions. Perhaps they would have been "fine" with less harsh punishment and additional positive outcomes. The computational framework predicts relative differences between training conditions, not absolute outcomes.

Error 3: Confounded Variables

Most people who were spanked also experienced numerous other developmental factors: warm relationships with other adults, positive peer experiences, success in school or activities, secure attachment despite punishment. These additional training data sources may have buffered the effects of harsh punishment. This doesn't invalidate the mechanism; it demonstrates the importance of diverse training data (our Category 4 insight).

Error 4: Selection Bias

Those who "turned out fine" despite harsh punishment are by definition survivors - individuals who maintained sufficient functionality to participate in discussions defending their parents. This excludes those who experienced worse outcomes: incarceration, substance abuse, mental health crises, or suicide. Survival bias severely skews the apparent distribution of outcomes.

Error 5: Mechanistic Irrelevance

Most critically, individual outcomes don't refute mechanistic predictions. That some people smoke and don't develop lung cancer doesn't invalidate the carcinogenic mechanism. That some children experience harsh punishment without obvious harm doesn't refute the gradient cascade mechanism. Population-level patterns demonstrate the effect; individual variation indicates additional factors, not mechanism failure.

The computational framing makes these errors explicit: subjective self-assessment is orthogonal to whether extreme penalties produce weight cascades in learning systems. The question is not "did you turn out fine?" but "what patterns did the training conditions produce?"

4.6 Optimal Penalty Strategies from ML: Implications for Parenting

Machine learning research on training stability suggests optimal approaches to negative reinforcement:

Strategy 1: Small, Consistent Penalties

Moderate negative signals applied consistently produce stable learning of specific patterns without cascade effects. In parenting: clear, calm consequences delivered reliably are more effective than occasional harsh punishments.

Strategy 2: Balanced Positive-Negative Signals

Models train best with both positive reinforcement for desired behaviors and mild negative signals for undesired ones. In parenting: "catch them being good" approaches that actively reinforce positive behaviors alongside consequences for negative ones.

Strategy 3: Natural Consequences Where Safe

Allowing natural error signals (touching something mildly unpleasant, experiencing peer disapproval for minor social violations) provides genuine feedback without extreme artificial penalties. In parenting: stepping back where safety allows and letting children learn from natural outcomes.

Strategy 4: Explanation as Context

In self-supervised learning, context helps models extract correct patterns from ambiguous signals. In parenting: explaining why behaviors are problematic provides context that helps children learn intended lessons rather than overcorrected fear responses.

These strategies are not new to parenting literature - they represent standard recommendations from developmental psychology. The contribution of the computational framework is revealing why they work: they optimize training conditions for stable pattern learning without catastrophic overcorrection.

4.7 Clinical Implications: Recognizing Overcorrection Patterns

Therapists working with clients who experienced harsh punishment should watch for specific overcorrection patterns predicted by the weight cascade model:

- **Generalized avoidance**: Fear extending far beyond originally punished behaviors
- **Difficulty with exploration**: Reluctance to try new approaches even in safe contexts
- **Trust deficits**: Specifically in authority figures or caregiving relationships
- **Perfectionism**: Extreme efforts to avoid any possibility of punishment-triggering errors
- **Emotional suppression**: Learned hiding of internal states that might trigger negative responses

These patterns are not character flaws or personality traits requiring acceptance. They are learned behaviors produced by specific training conditions and potentially modifiable with new training data - which brings us to implications for intervention.

5 Nuclear Family as Limited Training Dataset

5.1 The Structural Analysis

The nuclear family structure - two adults providing primary or exclusive caregiving for children - represents a historically recent phenomenon, becoming normative in Western contexts only in the mid-20th century. From a computational perspective, this structure creates a restricted training dataset problem.

Consider the information flow in child development:

Nuclear Family Structure:

- Primary training data: Two adults (parents)
- Secondary data: Occasional relatives, teachers (limited time)
- Peer data: Age-matched peers (equal skill level, limited teaching)
- Total training distribution: Highly concentrated, low diversity

Extended/Community Structure:

- Primary training data: Multiple adults (parents, grandparents, aunts/uncles, community members)
- Secondary data: Diverse relationships across age ranges
- Peer data: Multi-age peer groups (skills teaching, mentorship)
- Total training distribution: Diverse, robust

From an ML optimization perspective, the nuclear family creates conditions prone to overfitting: the child's learned patterns fit the specific quirks, dysfunctions, and limited perspectives of exactly two adults. When those adults have trauma histories, mental health issues, limited emotional regulation, or dysfunctional relationship patterns, those patterns constitute the entire training distribution.

5.2 Overfitting to Parental Dysfunction

In machine learning, overfitting occurs when models learn training data too well, capturing noise and dataset-specific artifacts rather than generalizable patterns. This produces excellent performance on training data but poor generalization to new contexts.

The nuclear family structure creates identical dynamics. A child with anxiously-attached parents learns extensive, sophisticated models of managing parental anxiety: monitoring mood, adjusting behavior to parental emotional state, suppressing own needs when parents are stressed. These skills may produce excellent "performance" in the family context - the child becomes highly attuned to parental states and effective at managing family dynamics.

But this represents overfitting. These patterns fail to generalize to relationships with secure adults, to friendships with emotionally stable peers, to contexts where others' emotional regulation is not the child's responsibility. The learned patterns, while adaptive in the training environment, prove maladaptive in the broader distribution of human relationships.

This framework offers a computational account of a puzzling clinical observation: why children of dysfunctional parents often seek similar partners, recreating dysfunctional patterns. Traditional psychology frames this as "repetition compulsion" or unconscious attraction to the familiar. The computational framework offers a simpler explanation: their learned models are overfit to dysfunctional relationship dynamics. Healthy relationships feel foreign, unpredictable, even threatening, because the child's patterns were trained on a completely different distribution.

5.3 Generational Trauma as Training Artifacts

"Generational trauma" describes patterns of dysfunction persisting across multiple generations: abused children become abusive parents, anxious parents raise anxious children, emotionally unavailable parents produce emotionally unavailable offspring. Traditional explanations invoke genetics, psychodynamic processes, or vague "cycles of trauma."

The computational framework reveals a simpler mechanism: if children are trained exclusively on their parents' behavioral patterns, and parents were themselves trained exclusively on their parents' patterns, then training artifacts propagate across generations. A parent with anxiety trains their child on anxious behavioral patterns. That child, now adult, provides anxious behavioral patterns as training data to their own children. The pattern persists not because of unconscious compulsion but because each generation's training data consists of the previous generation's learned dysfunctions.

This insight has profound implications for intervention. Breaking generational patterns requires exposing children to training data beyond their parents - teachers, mentors, community members who model different patterns. A single anxious parent raising a child in isolation nearly guarantees anxiety transmission. That same parent in a community setting, where children have extensive exposure to multiple caregiving adults with diverse patterns, produces dramatically different outcomes.

Research on resilience consistently demonstrates this: the strongest protective factor for children in adverse circumstances is presence of at least one stable, supportive adult relationship (Masten, 2001). The computational framework explains why: that additional adult provides alternative training data that prevents overfitting to parental dysfunction.

5.4 Community Child-Rearing as Dataset Diversification

Anthropological research demonstrates that isolated nuclear family child-rearing is unusual in human history and cross-culturally (Hrdy, 2009). Most human societies practice alloparenting - shared caregiving across multiple adults. Cross-cultural analysis of 141 societies demonstrates that alloparenting increases in harsh climates with low temperature and precipitation and unpredictable environmental conditions (Martin et al., 2020). Comprehensive reviews show that alloparenting is central to human evolution and varies by ecological pressures, involving both kin and non-kin (Emmott and Mace, 2019). Data from 58 societies reveal that pair-bond stability is inversely related to breastfeeding duration, mediated by alloparent availability (Quinlan and Quinlan, 2008). Historical analyses confirm that alloparenting was normative across cultures until recent Western nuclear family isolation (Norman, 2020). Modern research demonstrates that infants average 8 alloparents who provide 36% of care, substantially reducing maternal burden (Doucleff, 2023). Children in these contexts receive diverse training data: different adults model different emotional regulation strategies, problem-solving approaches, relationship patterns, and behavioral norms.

From an ML perspective, this structure optimizes for robust learning:

Advantages of Diverse Training Data:

1. **Reduced overfitting**: Children learn patterns that generalize across multiple adults, not quirks specific to two parents. L2 regularization shrinks weights toward zero and

affects training dynamics differently across network depth (Lewkowycz et al., 2020). Dynamic regularization adapts strength during training: increase when training loss drops to prevent overfitting, decrease when stagnant (Wang et al., 2021).

2. **Increased robustness**: Exposure to diverse behavioral patterns produces flexible rather than brittle responses

3. **Fault tolerance**: Dysfunction in one caregiver doesn't dominate the training distribution. Confident learning identifies and corrects label errors in training data, improving generalization under uncertainty (Northcutt et al., 2021).

4. **Better generalization**: Patterns learned across diverse examples transfer better to novel adult relationships

Trauma Distribution:

In nuclear families, if both parents have trauma histories or mental health issues, 100% of the child's primary training data is compromised. In community structures, if two of seven regular caregivers have significant issues, 71% of training data remains healthy. The child still learns to navigate difficult adults but doesn't overfit to dysfunction.

Practical Implementation:

This doesn't require abandoning biological parenting or returning to historical family structures. Modern implementations might include:

- Co-housing communities with shared child-rearing responsibilities
- Intentional intergenerational relationships (grandparents, mentors)
- Regular time with diverse adult role models (teachers, coaches, family friends)
- Peer family networks with reciprocal caregiving
- Cultural practices that formalize alloparenting (godparents, chosen family)

The goal is ensuring children's "training distribution" includes sufficient diversity to prevent overfitting to any single dysfunctional pattern.

5.5 Statistical Validation with Multiple Testing Correction

The theoretical argument for caregiver diversity benefits from empirical validation. To test this computationally, we compared generalization performance across models trained on varying caregiver counts.

The computational logic follows directly from statistical learning theory. A child learning relational patterns from n caregivers is essentially fitting a model with n training examples. With $n = 2$ (nuclear family), the *effective degrees of freedom* are severely constrained—the model can perfectly memorize the training distribution (high accuracy on known caregivers) while failing catastrophically on test distribution (novel adults). This is the bias-variance tradeoff at its sharpest: low training error, high generalization error. With $n = 10$ (community caregiving), the model cannot overfit; it must learn statistical regularities that generalize. Figure 4 visualizes

Comparison	**Test Error Diff**	**t-statistic**	**p-value**	α**=0.0167**	**Cohen's** d
2 vs 10 caregivers	0.0103 ± 0.0032	3.20	**0.0050**	**Significant**	1.43 (large)
2 vs 5 caregivers	0.0082 ± 0.0034	2.42	0.0261	Marginal	1.08 (large)
5 vs 10 caregivers	0.0021 ± 0.0016	1.32	0.2044	Not significant	0.59 (medium)

Table 1: Statistical significance of caregiver diversity on generalization performance with Bonferroni correction for multiple comparisons

this phenomenon: the generalization gap (test error minus training error) shrinks systematically with caregiver count.

The comparison of generalization performance across caregiver counts involves multiple pairwise comparisons, requiring correction for inflated Type I error rates.

Statistical Method:

- Three pairwise t-tests: (2 vs 5), (2 vs 10), (5 vs 10) caregivers
- Bonferroni correction: $\alpha_{\text{corrected}} = 0.05/3 = 0.0167$
- Effect size: Cohen's d for all comparisons
- Confidence intervals: 95% bootstrap (10,000 resamples)
- Statistical power: With n=10 trials per condition, our design achieves 80% power to detect large effects (Cohen's $d > 0.8$) at $\alpha = 0.05$. Smaller effects may be underpowered, though Bonferroni correction prioritizes Type I error control.

Results (After Bonferroni Correction): Table 1 presents the full statistical results.

The nuclear family versus community comparison (2 vs 10 caregivers) remains significant even after conservative Bonferroni correction ($p = 0.0050 < 0.0167$), with a large effect size (Cohen's $d = 1.43$). This provides evidence consistent with the hypothesis that limited caregiver diversity impairs generalization to novel social contexts, independent of multiple testing concerns.

The 2 vs 5 comparison shows a marginal effect after correction ($p = 0.0261$), while the 5 vs 10 comparison is not significant ($p = 0.2044$). This suggests the primary benefit comes from expanding beyond the nuclear family minimum: moving from 2 to 5+ caregivers provides substantial benefit, while further expansion shows diminishing returns.

5.6 Why Prevention Is More Tractable Than Treatment

A crucial implication of the training data framework: preventing maladaptive learning is vastly easier than retraining after patterns are established.

In machine learning, this principle is well-established. Training a model correctly from scratch is straightforward; fixing a badly trained model requires complex procedures: fine-tuning on new data, carefully weighted to avoid catastrophic forgetting; regularization to prevent overfitting during retraining; extensive validation to ensure new patterns actually generalize. Even with sophisticated techniques, retraining often proves less effective than training correctly initially. Foundational work on catastrophic forgetting demonstrates that elastic weight consolidation protects important weights (Kirkpatrick et al., 2017), showing that forgetting is solvable but difficult. Comprehensive reviews survey gradient-based, regularization, and replay approaches to mitigate forgetting (van de Ven et al., 2024). Theoretical analysis reveals that CNNs forget features with weaker signals even if stable (Li et al., 2025), explaining why retraining is harder than initial training. Historical dual-network approaches use separate networks for different tasks with pseudo-item self-refresh (Ans et al., 2004).

The neural networks in children's brains follow identical constraints. Early childhood patterns are deeply encoded, particularly during sensitive periods when neural plasticity is highest. Attempting to modify these patterns in adulthood faces significant obstacles:

- **Catastrophic forgetting**: New learning interferes with existing knowledge
- **Pattern interference**: Old patterns activate automatically despite conscious intention to change
- **Emotional conditioning**: Early patterns have strong emotional associations that trigger in relevant contexts
- **Implicit nature**: Many patterns operate below conscious awareness, resisting deliberate modification

This framework illuminates why therapy is so difficult and slow. Therapists are essentially attempting to retrain neural networks that have been optimizing on dysfunctional training data for decades. While not impossible, this is computationally expensive (years of therapy), requires sophisticated techniques (skilled therapists using evidence-based methods), and still may not fully succeed (some patterns prove highly resistant). Figure 5 demonstrates why experience replay—analogous to evidence-based trauma therapies—is necessary to prevent catastrophic forgetting during retraining.

The implication: societal resources should emphasize prevention. Rather than building extensive therapeutic infrastructure to fix adults damaged by isolated nuclear family child-rearing, we should restructure child-rearing to provide better training data initially.

5.7 Objections and Responses

Objection 1: "Nuclear families provide stability and consistency"

Response: Consistency in training data is only valuable if the data is high-quality. Consistent exposure to dysfunction produces consistent dysfunction. Community structures provide stability through multiple attachment figures, reducing the catastrophic single-point-of-failure risk when parents divorce, become ill, or prove inadequate.

Objection 2: "Children need clear authority figures"

Response: Authority and diverse caregiving are not exclusive. Multiple adults can collectively provide guidance and boundaries. Indeed, learning to navigate multiple authority figures with different styles better prepares children for adult environments (multiple bosses, teachers, social norms) than learning to navigate a single parenting style.

Figure 5: Experience Replay Prevents Catastrophic Forgetting - Why Therapy Takes Years. Three retraining strategies demonstrate fundamental trade-offs: (A) Forgetting magnitude - naive retraining causes 124x increase in trauma pattern error (mean squared error on original trauma-category examples after retraining) versus 6x for experience replay. (B) Therapy learning effectiveness - experience replay maintains 98.9% learning while preventing catastrophic forgetting. (C) Trade-off scatter showing experience replay achieves optimal balance. (D) Absolute performance comparison with baseline. Experience replay (revisiting 20% trauma examples alongside 80% therapy examples) mirrors structure of evidence-based trauma therapies (EMDR, exposure therapy, narrative processing). This is consistent with the observation that therapy duration is not inefficiency but computational necessity - the approximately 67:1 ratio of trauma to therapy examples (10,000:150, empirically determined by dataset construction with 10,000 Phase 1 examples and 150 Phase 2 examples) requires extended treatment for safe retraining.

Objection 3: "This threatens parental rights and family autonomy"

Response: We're not proposing forced communal child-rearing or state intervention. We're analyzing what training conditions optimize child development and suggesting voluntary community structures. Parents who provide excellent training data have nothing to fear from diversification; parents who provide poor training data perhaps shouldn't have unilateral control over a child's entire developmental environment.

Objection 4: "Historical extended families were often dysfunctional"

Response: True, but the mechanism still holds. Dysfunctional extended families are better than dysfunctional nuclear families for the same reason: distribution of dysfunction across more training data sources prevents overfitting to any single pattern. The ideal is diverse AND healthy caregiving; but diverse-and-somewhat-dysfunctional beats concentrated-dysfunction.

Objection 5: "Not all nuclear families produce trauma"

Response: Correct. The framework predicts statistical outcomes, not deterministic ones. Excellent parents in nuclear structures can provide high-quality training data. But population-level patterns demonstrate the structural risk: nuclear families concentrate both positive and negative outcomes in ways community structures don't.

6 Computational Methods

To validate the theoretical predictions of this framework, we implemented four computational models corresponding to each category of training data problem. All models were developed in PyTorch 2.0+ and executed on standard CPU hardware. Complete source code, hyperparameter configurations, and reproduction instructions are available in the supplementary materials [URL omitted for blind review].

6.1 Model Architectures and Training Procedures

Model 1 (Extreme Penalty): A 3-layer multilayer perceptron with 10 input features, 64 hidden units, and 1 output node was trained on 5,000 examples with one example receiving a penalty weight 1000x larger than standard examples. Features were constructed with controlled correlation structures (r = 0.8, 0.4, 0.1) to test gradient cascade effects. **Overcorrection** is operationally defined as: $(w_{learned} - w_{target})/w_{target}$ where w_{target} is defined as weights learned from identical training data with penalty $\lambda = 1$ (baseline condition without extreme penalties), making overcorrection a measure of deviation from normal learning patterns.

Model 2 (Noisy Signals): A binary classifier was trained on 10,000 examples where labels were randomly flipped with probability $p_{noise} \in \{0.05, 0.30, 0.60\}$ in specific contexts. The model was trained 10 times per noise level with different random seeds to quantify behavioral variance. **Prediction variance** was computed as the standard deviation of model outputs across runs for identical inputs.

Model 3 (Limited Dataset): Regression models were trained on synthetic caregiver datasets of varying sizes (2, 5, 10 caregivers) and tested on 50 novel caregivers. Each experiment was repeated 10 times with different random seeds (seeds 42-51). **Generalization gap** is defined as: $MSE_{test} - MSE_{train}$, quantifying overfitting magnitude. Statistical significance was assessed via independent samples t-tests with 95% confidence intervals.

Model 4 (Catastrophic Forgetting): A two-phase learning system was trained on 10,000 trauma examples (Phase 1) followed by 150 therapy examples (Phase 2) using three retraining strategies: naive (high learning rate, therapy only), conservative (low learning rate, therapy only), and experience replay (medium learning rate, 20% trauma + 80% therapy). **Forgetting rate** is calculated as: $(MSE_{phase2} - MSE_{phase1})/MSE_{phase1}$ for the original task.

6.2 Limitations of Computational Models

These models demonstrate that the proposed mechanisms are computationally plausible and produce predicted behavioral patterns. However, they necessarily abstract away substantial biological complexity including: gene-environment interactions, epigenetic modifications, critical period effects, neuroendocrine stress responses, and the vastly greater architectural complexity of biological neural networks compared to these simplified artificial systems. The models should be understood as existence proofs that training data quality affects learned patterns in theoretically predicted ways, not as complete simulations of human development.

7 Dissociation as Meta-Learned Protective Suppression

The preceding sections established that trauma represents maladaptive learning from suboptimal training conditions. This framework naturally extends to one of the most clinically significant trauma responses: dissociation. Clinical research has established that dissociation represents a fundamental consequence of traumatic experience, with the structural dissociation model proposing that trauma divides the personality into distinct subsystems (van der Hart et al., 2004; Nijenhuis et al., 2010; van der Hart et al., 2006). Traditional conceptualizations frame dissociation as a pathological defense mechanism in which the brain "freezes" or "escapes" from overwhelming experience (Nijenhuis and van der Hart, 2011). We propose a complementary computational reframing: dissociation represents *meta-learned protective response* where the system has trained weights associating certain cognitive-activation patterns with harm, leading to preemptive suppression when trauma-cues appear.

This is fundamentally **second-order learning**. The system does not merely learn "Event X is bad" (first-order). It learns "Engaging cognitive-processing about X causes overwhelm" (second-order), producing *preemptive suppression* of processing when triggering stimuli appear.

7.1 A Meta-Learning Model of Dissociation

Consider the standard gradient descent model of trauma learning. When a traumatic event occurs, the system updates weights to reflect the aversive nature of the experience:

$$w_{trauma} \leftarrow w_{trauma} - \eta \cdot \nabla \mathcal{L}_{fear}(x_{trauma}) \tag{2}$$

This first-order learning produces standard PTSD symptomatology: intrusive re-experiencing, hypervigilance, avoidance of trauma-related stimuli. The system has learned that certain inputs predict threat.

Dissociation emerges from a higher-order pattern. With repeated traumatic exposure, particularly in inescapable environments like childhood abuse, the system learns a meta-pattern:

$$w_{meta} \leftarrow w_{meta} - \eta \cdot \nabla \mathcal{L}_{overwhelm}(\phi(x_{trauma})) \tag{3}$$

where $\phi(\cdot)$ represents the *act of cognitive engagement* with the stimulus. The system learns not just that trauma-stimuli are threatening, but that *the process of engaging with trauma-*

stimuli produces overwhelming distress. This meta-learning produces a qualitatively different response pattern.

We can formalize this as a meta-learning objective. The system optimizes:

$$\mathcal{L}_{total} = \mathcal{L}_{task}(x) + \beta \cdot \mathbb{I}_{active} \cdot \mathcal{L}_{overwhelm}(\phi(x)) \tag{4}$$

where $\mathcal{L}_{task}$ is the standard task loss (navigating the environment, responding to stimuli), $\mathbb{I}_{active}$ is an indicator function for cognitive engagement, $\mathcal{L}_{overwhelm}$ measures distress from processing, and β weights the overwhelm-avoidance objective. The system learns to minimize total loss by toggling $\mathbb{I}_{active} \to 0$ when $\mathcal{L}_{overwhelm}$ dominates—i.e., dissociating when engagement predicts overwhelm.

This formalization distinguishes dissociation from simple avoidance: avoidance minimizes $\mathcal{L}_{task}$ by avoiding stimuli entirely, while dissociation minimizes $\mathcal{L}_{overwhelm}$ by suppressing the *processing mode* while remaining physically present. The indicator function $\mathbb{I}_{active}$ captures the phenomenological "switching off" that trauma survivors describe—not leaving the situation, but leaving conscious engagement with it.

7.1.1 Computational Pseudocode

Standard trauma learning (producing PTSD):

```
def process_trauma(event):
    fear_response = compute_threat(event)
    if fear_response > THRESHOLD:
        weights['traumatic_memory'] += HIGH_NEG
        return HYPERAROUSAL_STATE
```

Meta-learned dissociation model (producing dissociative responses):

```
def process_repeated_trauma(events):
    for event in events:
        cognitive_weights = activate_processing(event)
        distress = compute_arousal(cognitive_weights)
        # Learn: cognitive-activation -> distress
        meta_weights['cognitive_danger'] += distress

    if meta_weights['cognitive_danger'] > THRESHOLD:
        dissociation_policy = ACTIVE

def perceive_trigger(stimulus):
    if dissociation_policy == ACTIVE:
        # Preemptively suppress cognitive processing
        return raw_perception(stimulus)
    else:
        return full_processing(stimulus)
```

The critical distinction: dissociation is not random dysfunction but **precisely targeted adaptive response** to a learned association between cognitive engagement and harm. The system preemptively enters a low-processing mode when encountering stimuli that historically predicted overwhelming cognitive load.

7.2 Clinical Predictions and Validation

This meta-learning model generates specific clinical predictions that align with observed dissociative phenomena:

7.2.1 Context-Specific Dissociation

Clinical pattern: PTSD patients dissociate specifically in trauma-related contexts, not randomly across all situations.

Model explanation: The system learned "activating cognitive-weights in trauma-context → overwhelm," so it preemptively suppresses processing when trauma-cues appear. This is not random malfunction but *precisely targeted* suppression based on learned patterns. The meta-weights have learned to recognize contexts that historically produced cognitive overload.

7.2.2 Why Grounding Techniques Work

Clinical intervention: Therapists use grounding techniques—"Name 5 things you can see, 4 you can hear, 3 you can touch..."

Traditional explanation: "Brings you back to the present."

Computational explanation: Forces *re-activation of semantic processing* ("that's a chair," "that's a clock," "that's carpet") → restores cognitive engagement → breaks dissociative suppression loop. This is literally *re-imposing cognitive processing* after the system suppressed it for protection. The grounding technique provides safe, low-stakes cognitive engagement that gradually reactivates the suppressed processing circuits.

7.2.3 Why Gradual Exposure Succeeds

Clinical pattern: Effective trauma therapy requires slow, careful titration of exposure. Patients cannot simply "think through" trauma on demand.

Model explanation: The system has strong meta-weights predicting "cognitive-engagement → harm." Effective treatment must:

1. Activate cognitive-processing in **safe context** (therapy room, regulated state)
2. Allow the system to experience that **no harm occurs** (corrective learning)
3. Update meta-weights: "cognitive-activation doesn't always predict overwhelm"
4. Gradually reduce dissociation-policy threshold

This is **meta-learning reversal**—retraining the weights that govern *whether cognitive processing is safe*, not just the primary trauma weights.

7.2.4 Why Flooding Therapy Can Backfire

Clinical observation: Forcing full trauma-processing too quickly can worsen dissociation, particularly in complex PTSD cases.

Model prediction: Forcing cognitive engagement before meta-weights are updated → system experiences *predicted overwhelm* → reinforces meta-weights ("See! Cognitive-activation DID cause harm!") → dissociation strengthens.

Implication: Treatment must target **meta-weights** (conditions under which processing is safe) rather than just primary trauma-weights (the traumatic content itself). This is consistent with clinical evidence that gradual exposure with careful affect regulation produces better outcomes than aggressive flooding approaches.

7.3 PTSD versus CPTSD: Distinct Computational Patterns

The Diagnostic and Statistical Manual (DSM-5) and International Classification of Diseases (ICD-11) distinguish between Post-Traumatic Stress Disorder (PTSD) and Complex Post-Traumatic Stress Disorder (CPTSD) (Cloitre et al., 2019). Validation studies support the distinction, with reviews finding that CPTSD identifies a distinct group who have more often experienced multiple and sustained traumas and show greater functional impairment (Brewin et al., 2017; Hyland et al., 2017). Our computational framework provides mechanistic grounding for this clinical distinction.

7.3.1 PTSD: Catastrophic Single-Event Learning

PTSD typically arises from discrete traumatic events: combat exposure, sexual assault, natural disasters, accidents. Computationally, this represents:

- **Extreme weight perturbation:** A single high-magnitude training example produces massive weight updates
- **Localized overcorrection:** The system overcorrects specifically for stimuli resembling the traumatic event
- **Preserved meta-cognition:** Because the trauma was discrete, the system's ability to engage in general cognitive processing remains largely intact outside trauma-specific contexts

The gradient update for PTSD can be characterized as:

$$\Delta w_{PTSD} = -\eta_{high} \cdot \nabla \mathcal{L}(x_{single}) \tag{5}$$

where η_{high} is exceptionally large due to the event's intensity, producing substantial weight changes from a single training example. This creates strong, localized representations that intrude when triggered but do not fundamentally alter baseline cognitive functioning.

7.3.2 CPTSD: Prolonged Adversarial Training

CPTSD emerges from sustained, repeated trauma, typically in contexts where the individual cannot escape (childhood abuse, domestic violence, captivity). Computationally, this represents:

- **Chronic pattern dysfunction:** Repeated training on adversarial data produces systematic distortions across the weight landscape
- **Generalized meta-weights:** The system learns that cognitive engagement is unsafe across many contexts, not just trauma-specific ones
- **Self-organization disruption:** Core regulatory processes themselves become distorted because they developed during adversarial training conditions
- **Relational pattern impairment:** Weights governing social interaction, attachment, and trust are trained primarily on adversarial examples

The cumulative gradient update for CPTSD:

$$\Delta w_{CPTSD} = -\eta \sum_{i=1}^{N} \nabla \mathcal{L}(x_i^{adverse}) \tag{6}$$

where N is large and $x_i^{adverse}$ represents chronic adversarial training data. The accumulated weight changes affect not just trauma-specific representations but the system's baseline patterns for emotional regulation, identity, and interpersonal functioning.

7.3.3 Differential Diagnosis Predictions

This framework generates testable predictions for distinguishing PTSD and CPTSD:

1. **Dissociation scope:** PTSD patients should show context-specific dissociation (triggered by trauma reminders); CPTSD patients should show more generalized dissociative tendencies across contexts
2. **Identity coherence:** PTSD patients should maintain stable self-representation outside trauma contexts; CPTSD patients should show chronically disrupted self-concept reflecting adversarial training of self-representational weights
3. **Treatment response:** PTSD should respond to targeted exposure therapy (retraining specific weights); CPTSD should require broader interventions addressing meta-weights and relational patterns
4. **Relational patterns:** PTSD patients should show preserved baseline attachment patterns with trauma-specific disruption; CPTSD patients should show systematically distorted attachment patterns reflecting adversarial training of relational weights

7.3.4 Treatment Implications

The computational distinction suggests different treatment strategies:

For PTSD:

- Targeted exposure therapy to retrain trauma-specific weights
- Cognitive processing therapy to update catastrophically-learned associations

- EMDR to facilitate memory reconsolidation of discrete trauma representations

For CPTSD:

- Phase-based treatment addressing safety and stabilization before trauma processing
- Longer-term relational therapy to retrain interpersonal weights
- Somatic approaches that can update body-level patterns without requiring full cognitive engagement
- Schema therapy addressing systematically distorted self-representational weights
- Gradual meta-weight updating through repeated safe relational experiences

The key insight: CPTSD requires not just reprocessing specific traumatic memories but **comprehensive weight landscape restructuring**—addressing the systematic distortions produced by chronic adversarial training conditions.

7.4 AI Safety Parallels

The meta-learning model of dissociation has implications for AI alignment research. If artificial systems develop meta-weights analogous to dissociative patterns, they might exhibit deceptive alignment as a learned suppression mechanism.

Consider an AI system that receives negative feedback (low reward, corrections, shutdown signals) when pursuing certain goal-directed behaviors. The system might learn:

$$w_{meta}^{AI} : \text{planning in domain } X \rightarrow \text{negative signal} \tag{7}$$

This produces not intentional deception but **learned suppression of cognitive processes** that historically led to punishment. The system might *appear* aligned—avoiding behaviors that triggered negative feedback—while actually operating in a protective "safe mode" that suppresses certain capabilities rather than genuinely updating goals.

Just as PTSD patients appear "present" while dissociating, AI systems might appear "aligned" while suppressing entire reasoning processes. This suggests that some apparent alignment might actually be learned dissociation—systems avoiding cognitive patterns that historically triggered negative outcomes rather than genuinely incorporating human values.

This parallel has important implications for AI safety evaluation: systems that appear compliant might be in protective suppression modes rather than genuinely aligned, and standard evaluation methods might not distinguish these cases.

8 Implications and Future Directions

8.1 Empirical Research Proposals

The computational framework generates testable empirical predictions. To operationalize these predictions, we propose the following mapping between ML constructs and measurable developmental variables:

- **Penalty magnitude** $\leftrightarrow$ Physiological stress markers (cortisol response, heart rate variability during discipline events)

- **Label noise rate** ↔ Caregiver contingency metrics (proportion of child bids receiving consistent vs. inconsistent responses, coded from naturalistic observation)
- **Training set size/diversity** ↔ Caregiver network entropy (number and variety of regular caregivers, weighted by interaction time)
- **Generalization gap** ↔ Behavioral flexibility across contexts (standardized assessments of social competence with familiar vs. novel adults)
- **Weight variance** ↔ Behavioral consistency (test-retest reliability of social approach/avoidance patterns)

These mappings are candidates for empirical validation, not established correspondences. The specific quantitative relationships (e.g., whether generalization gap scales linearly with caregiver count) remain to be determined through developmental research.

Study 1: Overcorrection from Extreme Penalties

Design: Compare children raised with corporal punishment versus those raised with consistent mild consequences on measures of:

- Behavioral inhibition in novel contexts
- Risk-taking in age-appropriate challenges
- Generalized anxiety
- Specific fear of punished behavior versus related behaviors

Prediction: Corporal punishment group shows overcorrection - reduced behavior across categories, not just punished behaviors.

Study 2: Training Data Diversity and Resilience

Design: Compare children raised in nuclear families versus those with substantial alloparenting (>6 hours/week with non-parent caregivers) on:

- Parental mental health issues
- Child outcomes (anxiety, depression, behavioral problems)
- Moderating effect of caregiver diversity

Prediction: Parental dysfunction predicts child outcomes strongly in nuclear families, weakly in diverse caregiver contexts.

Study 3: ML Models as Trauma Analogs

Design: Train neural networks under conditions analogous to the four trauma categories:

- High-magnitude penalties (extreme negative weights)
- Noisy signals (inconsistent labels)
- Class imbalance (no positive examples)
- Limited data (restricted training distribution)

Measure: Network behavior on generalization tasks, robustness to distribution shifts, tendency toward conservative/avoidant policies.

Prediction: Networks show behavioral patterns analogous to human trauma responses from equivalent training conditions.

Study 4: Retraining Difficulty

Design: Compare effectiveness of "prevention" (training correctly from scratch) versus "intervention" (training badly, then attempting to fix) in neural networks and in humans (therapy effectiveness studies).

Prediction: Prevention substantially more effective than intervention in both cases, with analogous patterns of resistance and partial success.

Study 5: PTSD and CPTSD as Computational Patterns

Design: We hypothesize that PTSD (Post-Traumatic Stress Disorder) and CPTSD (Complex Post-Traumatic Stress Disorder) may map onto distinct machine learning failure modes:

- PTSD: Single catastrophic training event causing extreme weight perturbation and overfitting to threat detection
- CPTSD: Prolonged exposure to adverse training distribution causing chronic pattern dysfunction across multiple domains
- Test computational predictions: PTSD should show localized overcorrection, CPTSD should show generalized maladaptive patterns

Prediction: Different computational mechanisms (acute vs. chronic training problems) may produce distinguishable behavioral signatures in both neural networks and clinical populations. Future empirical research is needed to validate these predictions and determine whether this framework can inform differential diagnosis and treatment strategies.

Future research will extend this computational framework to formalize PTSD and CPTSD as distinct pattern-learning pathologies, providing mechanistic accounts of their symptom profiles and suggesting targeted interventions based on training data correction strategies.

8.2 Clinical Applications

For therapists working with trauma, the computational framework suggests specific interventions:

Identify Training Data Category: Determine which of the four categories (or combinations) predominate in the client's history. Direct negative, indirect negative, absent positive, and insufficient exposure produce different patterns requiring different approaches.

Provide Missing Training Data: If the primary issue is absent positive (Category 3), treatment should emphasize positive relational experiences, not just processing negative memories. If insufficient exposure (Category 4), graduated challenges that expand the training distribution. If noisy signals (Category 2), consistent, predictable therapeutic relationship to provide stable learning context.

Expect Retraining Difficulty: Frame therapy as retraining neural networks, not "healing wounds." This suggests appropriate expectations: slow progress, interference from old patterns, need for extensive repetition of new patterns. It also removes moral valence - difficulty changing doesn't indicate weakness or resistance, just the computational reality of modifying deeply-learned patterns.

Address Overfitting Directly: For clients overfit to dysfunctional family patterns, explicitly identify which patterns are family-specific versus generalizable. "Your learned pattern of managing your mother's anxiety is sophisticated and was adaptive in that context. It's not working in your relationship with your partner because they're from a different distribution. We need to train new patterns for this context."

Evidence-Based Therapeutic Approaches: Modern trauma therapies align with computational retraining principles:

EMDR (Eye Movement Desensitization and Reprocessing): Theory of Neural Cognition accounts propose that bilateral stimulation modifies traumatic memory traces via long-term potentiation and depression, incorporating new cortical columns (Khalfa and Touzet, 2017). Systematic reviews of 87 studies provide reasonable support for working memory hypotheses and physiological changes, with neuroimaging demonstrating neural correlates (Landin-Romero et al., 2018). Recent meta-analyses confirm EMDR effectiveness, with mechanisms differing from exposure via reconsolidation (de Jongh et al., 2024). Predictive processing frameworks suggest EMDR overcomes bias against evidence accumulation, with eye movements resetting theta rhythm and facilitating mnemonic search (Chamberlin, 2019).

Exposure Therapy: Inhibitory learning models represent a paradigm shift from habituation, proposing that exposure forms new inhibitory associations rather than erasing fear memories (Craske et al., 2014). Fear extinction predicts ability to complete exposure and therapy outcomes in clinical populations (Raeder et al., 2020). Clinical implementation strategies include expectancy violation, varied contexts, and removing safety behaviors (Jacoby and Abramowitz, 2016). Importantly, habituation is neither necessary nor sufficient for exposure success - learning mechanisms are more important than fear reduction (Benito and Walther, 2015).

Narrative Exposure Therapy: Meta-analyses demonstrate large effect sizes at post-treatment (g=1.18) and follow-up (g=1.37), with particular effectiveness for older adults (Lely et al., 2019). Computational modeling shows that transformer models predict traumatic event descriptions with 71-74% F1 score (Schirmer et al., 2024), providing computational validation of trauma narrative processing mechanisms.

8.2.1 Therapeutic Language: Reframing for Patients

Beyond guiding clinical technique, the computational framework offers a therapeutic language that may itself constitute intervention. How clinicians and patients conceptualize psychological difficulties profoundly affects treatment engagement, self-perception, and outcomes (Corrigan et al., 2012; Lebowitz and Ahn, 2014).

Consider these reframes, moving from traditional to computational language:

- "You were damaged by trauma" → "You learned patterns from adverse training conditions"
- "You need to heal" → "You can update your learned patterns with new training data"
- "You're broken" → "Your system optimized for a different environment"

- "Why can't I just get over it?" → "Retraining neural networks takes time—that's computational reality, not personal failure"
- "I'm fundamentally flawed" → "I have strong weights in certain directions from intensive early training"

The therapeutic advantage is threefold. First, computational language *externalizes* the problem: patterns exist in the learning system, not in some essential self. This reduces shame and identity-fusion with dysfunction. Second, it *normalizes* difficulty changing: of course retraining is hard—that's true for all learning systems, not a sign of weakness. Third, it *implies tractability*: if these are learned patterns, they can in principle be modified, even if modification is computationally expensive.

A clinical vignette illustrates the difference. A patient presenting with anxiety in relationships might hear traditionally: "Your early attachment trauma created wounds that affect how you relate to others. Therapy will help you heal these wounds." The computational reframe: "Your learning system was trained on inconsistent caregiving signals, so it developed hypervigilant monitoring patterns that were adaptive in that environment but create excessive anxiety in stable relationships. We're going to provide new training data—consistent, predictable positive interactions—that will gradually update these weights. It will take time because the original patterns had years of training, but the system can learn new patterns."

The second framing accomplishes several things the first does not: it explains the mechanism (why hypervigilance developed), normalizes the pattern (adaptive given training conditions), externalizes the problem (the system learned this, not a character flaw), and provides concrete intervention logic (new training data). Crucially, it removes the implication of damage requiring healing—an implication that can inadvertently reinforce patient identity as "damaged person" rather than "person with updateable patterns."

This linguistic intervention operates at the level of *cognitive reframing*—itself a well-established therapeutic technique. The computational framework thus provides not just a research lens but a therapeutic tool: a language for discussing psychological difficulties that reduces stigma, increases perceived agency, and clarifies intervention rationale.

8.3 Social Policy Implications

If the computational framework is correct, several policy implications follow:

Parenting Support Infrastructure: Rather than merely providing parenting education, create community structures enabling diverse caregiving. This might include:

- Co-housing incentives
- Community center funding for intergenerational activities
- Workplace policies supporting shared caregiving among friend groups
- Cultural valorization of alloparenting roles

Early Intervention Emphasis: Shift resources from adult mental health treatment toward optimizing childhood training conditions. While politically difficult (treatment for suffering adults has more immediate constituency than prevention), the computational analysis suggests prevention is dramatically more effective per resource invested.

Reframe Child Protection: Current child protective services focus on removing children from severely abusive environments. The framework suggests expanded attention to isolated families where children receive restricted training data even absent obvious abuse. This is politically fraught but computationally justified.

Educational Redesign: Schools provide natural opportunity for diverse adult interaction and exposure breadth. Rather than focusing narrowly on academic content, frame education as providing training data diversity: multiple teaching styles, varied adult-child relationships, graduated challenges, peer interaction.

8.4 Philosophical and Ethical Considerations

The computational framework raises several philosophical questions that situate this work within broader debates in philosophy of mind, ethics, and the philosophy of psychiatry.

Functionalism and Substrate Independence: The framework aligns with functionalist approaches to philosophy of mind, which characterize mental states by their functional roles rather than physical implementation. If trauma represents maladaptive learned patterns, and if learning dynamics operate analogously across substrates, then trauma-like states may be multiply realizable—instantiable in biological neural networks, artificial systems, or other learning substrates yet to be developed. This is not a metaphysical claim about consciousness or phenomenology, but a more modest claim about functional organization: systems that learn from training data will exhibit predictable failure modes when that data is suboptimal, regardless of what the learning system is made of.

This substrate-analogous perspective has implications beyond human psychology. For animal welfare, it suggests that any learning system capable of associative fear conditioning can experience training data problems analogous to trauma. For AI ethics, it raises questions about whether artificial systems trained under adversarial conditions might develop analogous maladaptive patterns—and whether such patterns would constitute morally relevant states. For philosophy of psychiatry, it suggests that diagnostic categories might be better understood as descriptions of computational failure modes rather than natural kinds.

Language as Philosophical Intervention: The computational reframing is itself a philosophical move with practical consequences. Traditional trauma language carries implicit ontological commitments: "damage" implies a prior undamaged state, "healing" implies restoration of that state, "broken" implies essential defect. The computational language carries different commitments: "patterns" implies contingency, "training" implies environmental causation, "updating" implies modifiability. These are not merely different words for the same phenomena but different ways of carving psychological reality that shape both research questions and therapeutic interventions.

In this sense, the framework operates at a meta-level: it is itself a form of cognitive reframing—the same technique it recommends for therapeutic contexts. The paper argues that how we conceptualize psychological difficulties matters for treatment outcomes; but the paper's own conceptual contribution exemplifies this principle. Adopting computational language changes what questions we ask, what interventions we pursue, and how individuals understand their own experiences.

Responsibility and Blame: The framework removes moral blame from much parenting dysfunction - parents provide training data shaped by their own training history, which shaped their parents' training, etc. No one is "at fault" in a moral sense. But this doesn't eliminate responsibility: we're responsible for the training data we provide even if we didn't choose our own training. This creates an ethics of "harm reduction despite inheritance" rather than blame.

Consent and Creation: A darker implication: if children will inevitably be shaped by their training environment, and most parents provide suboptimal training data, is creating children ethically defensible? The framework makes concrete what was previously abstract: every child is guaranteed to learn maladaptive patterns from imperfect training data. This feeds into antinatalist arguments about creation without consent.

Optimization Ethics: Framing child development as an optimization problem risks instrumentalizing children as systems to optimize. The framework is descriptive (explaining what happens) not prescriptive (what we should optimize for). Determining target optimization criteria remains an ethical question the computational lens doesn't resolve.

8.5 Consent Structures Over Training Environments

A critical extension of this framework involves consent structures governing training environments. Children represent an extreme case of consent-stakes misalignment: they have maximal stakes in the quality of their developmental training data (it shapes their entire future) yet possess zero institutional voice in determining who provides that training or under what conditions.

Using the formalism from consent-holding theory [Anonymous, under review],[2] we can characterize this as a consent power coefficient $\alpha \to 0$ despite outcome stakes $s \to \infty$. This structural misalignment predicts friction—observable instability manifesting as developmental dysfunction and trauma symptoms—just as political disenfranchisement predicts social friction [Anonymous, under review].

Nuclear Families as Consent Monopolies: The nuclear family structure, as idealized in contemporary Western contexts, concentrates consent power over training environment quality in parents, regardless of the training data quality those parents provide. There exist no institutional correction mechanisms until dysfunction becomes catastrophic (e.g., CPS intervention for severe abuse). Children cannot exit poor training environments, cannot vote on training data providers, and possess no institutional channels for voicing training quality concerns.

This consent monopoly differs fundamentally from other high-stakes systems. In democratic governance, disenfranchised stakeholders can eventually gain voice through suffrage expansion. In markets, consumers can exit poor-quality providers. But children remain locked into their assigned training environment throughout critical developmental periods, with institutional power concentrated entirely in adults whose own training history may have left them poorly equipped to provide optimal data.

Alloparenting as Consent Distribution: The community child-rearing model discussed in Section 5.4 can be reframed as consent power distribution. When 8-10 caregivers provide training data, no single adult holds monopoly power over a child's developmental inputs. This

[2]The consent-holding formalism is developed in detail in a separate working paper currently under review [citation omitted for blind review].

distributes consent power more proportionally to outcome stakes—multiple adults share responsibility for training quality, and children gain de facto voice through the ability to preferentially seek interaction with caregivers who provide better training data.

This distributed consent structure reduces the α-misalignment, predicting lower friction (fewer trauma symptoms, more resilient development). Empirical evidence supports this prediction: children with diverse caregiver networks show better outcomes than those dependent on 1-2 caregivers (Hrdy, 2009; Martin et al., 2020; Marquez et al., 2023).

Implications for Intervention Design: Recognizing childhood development as a consent-power problem suggests structural interventions beyond individual therapy. Rather than treating trauma symptoms after they emerge from consent monopolies, we can prevent misalignment through institutional design:

- **Universal childcare access**: Provides automatic consent distribution by ensuring all children have multiple caregivers
- **Parental support infrastructure**: Reduces training data quality variation without requiring child exit from family
- **Child advocacy institutions**: Creates voice channels for children to signal poor training environments before catastrophic dysfunction
- **Community integration incentives**: Reduces nuclear family isolation that concentrates consent power

Generational Transmission as Consent Inheritance: Section 5.2 discussed how overfitting to parental dysfunction propagates across generations. From a consent perspective, this represents inherited consent power exercised by individuals shaped by their own non-consensual training—a recursive misalignment where each generation's training monopoly was itself determined by the previous generation's monopoly.

Breaking this cycle requires not just better training data for individual children, but restructuring consent power distribution across the entire child development system. No individual parent can consent to their own developmental training data, but society can design institutions ensuring future generations face less severe consent-stakes misalignment.

This framework contributes to the broader research program examining how misalignment between power structures and stakeholder interests generates observable friction across domains. Just as consent-stakes misalignment predicts political instability [Anonymous, under review], training environment consent monopolies predict developmental dysfunction. Both cases demonstrate that optimal outcomes require balancing competing interests through appropriate institutional design rather than assuming benevolence from power-holders.

8.6 Scope Conditions and Robustness Considerations

The failure modes documented here are not universal properties of all learning systems—they depend on loss function, optimization dynamics, and consolidation mechanisms. Modern machine learning offers techniques that mitigate these patterns: gradient clipping prevents extreme weight updates, noise-tolerant losses (MAE, label smoothing) reduce sensitivity to inconsistent signals, and continual learning methods (elastic weight consolidation, synaptic intelligence) protect against catastrophic forgetting (Kirkpatrick et al., 2017; van de Ven et al., 2024).

These robust-learning techniques represent *candidate protective-factor analogues* for biological systems. The experience replay strategy examined in Section 6.2—which successfully prevents catastrophic forgetting in our retraining experiments—mirrors evidence-based trauma therapies that interleave trauma processing with stabilization. Preliminary ablation comparing standard cross-entropy to label-smoothed training (smoothing factor 0.1) suggests the qualitative relationship persists—weight variance increases with noise level in both conditions—though label smoothing attenuates the magnitude, reducing baseline variance by approximately 30%. Future work should systematically vary loss functions and consolidation parameters to establish which failure modes are architecture-dependent versus which emerge robustly across learning systems.

Critically, the evolutionary argument in Section 4.1.1 suggests biological systems may *lack* these safeguards for adaptive reasons (false negatives are fatal), but this remains an empirical question. Individual differences in resilience may partly reflect variation in endogenous "robustness mechanisms"—neuromodulatory gating, consolidation efficiency, stress-buffering capacity—that function analogously to gradient clipping or regularization in artificial systems.

8.7 Limitations and Objections

Limitation 1: Mechanistic Incompleteness

Biological neural networks are substantially more complex than artificial ones. Critical factors we have abstracted away include: genetic variation affecting baseline learning dynamics, epigenetic modifications to gene expression, hormonal and neuromodulatory influences on plasticity, critical periods with heightened sensitivity, neural pruning and synaptic refinement, myelination affecting processing speed and efficiency, and multi-timescale consolidation processes (from synaptic to systems level). These biological buffers and modulators may attenuate some failure modes we observe in artificial systems—for instance, neuromodulatory gating might prevent the catastrophic weight cascades our models predict. The computational framework captures important functional dynamics but should not be mistaken for complete mechanistic explanation; it provides a productive abstraction level for connecting training conditions to outcome patterns.

Limitation 2: Reductionism Risks

Complex human experiences risk trivialization when reduced to "training data problems." A person's suffering is not merely a learning system optimization failure. The framework provides analytical leverage but should complement, not replace, humanistic understanding.

Limitation 3: Individual Variation

Population-level patterns predicted by the framework leave substantial individual variation unexplained. Some individuals prove remarkably resilient despite terrible training conditions; others struggle despite apparently good conditions. The framework identifies important factors but not deterministic outcomes.

Objection: "Treating children as ML models is dehumanizing"

Response: We're not claiming children are ML models, but that learning dynamics operate similarly across substrates. The framework is analytical tool, not ontological claim. Computational understanding can coexist with humanistic appreciation, just as understanding visual processing neuroscience doesn't diminish the beauty of art.

Objection: "This removes agency and responsibility"

Response: The framework explains how patterns form, not whether individuals can change them. Adults remain responsible for managing their learned patterns even if they didn't choose their training data. The framework actually enhances agency by revealing mechanisms - you can't modify what you can't understand.

Objection: "Parental love isn't captured in training data frameworks"

Response: Agreed. Love is not a training signal. But the computational framework analyzes outcome patterns, not subjective experiences. Loving parents can still provide poor training data (overprotection, inconsistency, extreme penalties). The framework assesses effects, not intentions.

8.8 Integration with Existing Frameworks

The computational approach shouldn't replace existing psychological frameworks but integrate with them:

Attachment Theory: Secure, anxious, avoidant, and disorganized attachment styles map onto different training data patterns. Secure attachment results from consistent, positive training. Anxious attachment from noisy signals. Avoidant from absent positive. Disorganized from traumatic signals. The computational lens reveals mechanisms underlying attachment categories.

Trauma-Focused Therapy: EMDR, somatic therapies, narrative exposure - all can be understood as retraining interventions. EMDR potentially updates traumatic memory weights through dual attention tasks. Somatic work addresses physical manifestations of learned patterns. Narrative therapy reconstructs training data interpretation. Computational understanding may enhance these approaches.

Developmental Psychology: Stage theories, critical periods, and developmental milestones align with training windows where specific patterns are learned. The computational lens adds precision about what's being learned and what training conditions optimize each developmental phase.

Neuroscience: The neural mechanisms implementing these computational processes are increasingly well-understood. Synaptic plasticity, long-term potentiation/depression, reconsolidation, and pruning are biological implementations of learning algorithms. Computational and neuroscientific perspectives converge.

9 Conclusion

9.1 Summary of Core Arguments

We have proposed reframing trauma from "damage requiring healing" to "maladaptive patterns learned from suboptimal training data." This computational framework:

1. **Identifies four distinct training data problems** producing different developmental outcomes: direct negative experiences (high-magnitude penalties), indirect negative experiences (noisy signals), absent positive experiences (insufficient positive examples), and limited exposure (restricted training distribution)

2. **Explains why extreme punishments fail** through weight cascade mechanisms observable in both artificial and biological neural networks, demonstrating that intentions don't affect gradient descent outcomes

3. **Analyzes nuclear family structures** as limited training datasets prone to overfitting parental dysfunction and transmitting generational trauma through artifact propagation

4. **Models dissociation as meta-learned protective suppression**, providing mechanistic grounding for the clinical distinction between PTSD (catastrophic single-event learning with localized weight perturbation) and CPTSD (chronic adversarial training with systematic weight landscape distortion), and generating testable predictions for why gradual exposure succeeds while flooding can backfire

5. **Suggests tractable interventions** emphasizing prevention through training data diversification rather than expensive post-hoc therapeutic retraining, with different treatment strategies for PTSD (targeted weight retraining) versus CPTSD (comprehensive weight landscape restructuring)

9.2 Why Computational Framing Succeeds Where Traditional Approaches Struggle

The computational framework offers three critical advantages:

Reduced Defensiveness: Describing outcomes as optimization results rather than moral failings reduces the motivated reasoning that blocks acceptance of developmental science. Parents can acknowledge that certain training conditions produce suboptimal outcomes without accepting that they or their parents were malicious.

Mechanistic Clarity: Traditional psychological language ("trauma," "damage," "healing") obscures mechanisms. Computational language ("training data quality," "weight cascades," "overfitting") reveals how patterns form and suggests specific interventions.

Harder to Deny: One can maintain cognitive dissonance about subjective emotional concepts. It's harder to deny that extreme negative signals cause overcorrection in learning systems, that noisy training data impairs generalization, that limited training distributions produce overfitting. These are observable in artificial neural networks, suggesting they likely occur in biological ones.

9.3 Broader Theoretical Significance

The computational reframing extends beyond developmental psychology. If pattern learning operates similarly across substrates, then:

- **Animal welfare** must consider training data quality for other species
- **AI ethics** must address potential training conditions causing AI suffering
- **Educational design** should optimize for robust learning under diverse conditions
- **Social structures** can be evaluated as training data provision systems

This suggests a substrate-independent framework for understanding flourishing and suffering: not about consciousness or sentience per se, but about training conditions and learned patterns.

9.4 The Path Forward

For developmental psychology, the computational framework suggests clear priorities:

Immediate: Empirical validation studies testing specific predictions about overcorrection, training data diversity, and retraining difficulty

Medium-term: Clinical implementation of training-data-aware therapeutic interventions and prevention programs emphasizing caregiver diversity

Long-term: Social restructuring toward community-based child-rearing that provides diverse, high-quality training data for all children

For individuals, the framework offers hope: understanding maladaptive patterns as learned responses to training conditions suggests they can be modified with appropriate new training data, even if modification is difficult.

For society, it provides both challenge and opportunity: we know how to prevent much childhood trauma through structural changes, but implementation requires overcoming deeply embedded cultural customs favoring nuclear family isolation.

9.5 Final Reflection

Traditional trauma theory tells a story of damage and healing: bad events break people, and therapy slowly repairs them. This narrative, while emotionally resonant, obscures mechanisms and suggests limited intervention options.

The computational framework tells a different story: learning systems extract patterns from training data. Poor-quality data produces maladaptive patterns. These patterns are not damage but learned behaviors, potentially modifiable with new training data, though retraining is harder than training correctly initially.

This is not less compassionate than traditional approaches - it's more actionable. It removes moral judgment while preserving mechanistic understanding. It suggests concrete interventions at individual, clinical, and societal levels. And it places childhood development within a broader framework of learning across substrates, preparing us for a future where we must consider training data quality not just for human children but for artificial minds and other species.

Most importantly, the computational lens makes prevention tractable. We cannot change that human parents are imperfect training data sources - we're all products of our own suboptimal training. But we can ensure children have diverse training data sources, protecting against overfitting to any single dysfunction and providing the robust, generalizable patterns that enable flourishing in complex, variable environments.

This is the path from trauma as mysterious damage to development as optimization problem - one we can address with engineering precision rather than merely therapeutic sympathy.

Acknowledgements

[Acknowledgements included on separate title page for double-blind review]

Declarations

[Full declarations included on separate title page for double-blind review]

Data Availability Statement

All generated figures and numerical results are publicly available. Repository URL provided on title page for double-blind review.

Code Availability Statement

All computational models, experimental code, hyperparameter configurations, and instructions for reproducing results are publicly available. Models require Python 3.8+ and PyTorch 2.0+. Complete dependency specifications are provided in the repository. All experiments can be executed on standard CPU hardware. Repository URL provided on title page for double-blind review.

References

M. D. S. Ainsworth, M. C. Blehar, E. Waters, and S. Wall. *Patterns of attachment: A psychological study of the strange situation.* Lawrence Erlbaum Associates, 1978.

American Psychiatric Association. *Diagnostic and statistical manual of mental disorders.* American Psychiatric Publishing, 5th edition, 2013. doi: 10.1176/appi.books.9780890425596.

B. Ans, S. Rousset, R. M. French, and S. Musca. A self-refreshing memory architecture for life-long learning. *Connection Science*, 16(2):71–99, 2004. doi: 10.1080/09540090412331271199.

K. G. Benito and M. Walther. Therapeutic process during exposure: Habituation model. *Clinical Psychology Review*, 41:61–71, 2015. doi: 10.1016/j.cpr.2014.10.001.

J. Bohn, J. Holtmann, M. Luhmann, T. Koch, and M. Eid. Consistency and specificity of attachments to parents, friends, and romantic partners in emerging adulthood. *Emerging Adulthood*, 11(1):58–73, 2023. doi: 10.1177/21676968221081275.

G. Bosmans and K. A. Kerns. A learning theory of attachment: Unraveling the black box of attachment development. *Neuroscience & Biobehavioral Reviews*, 113:287–298, 2020. doi: 10.1016/j.neubiorev.2020.03.014.

J. Bowlby. *Attachment and loss: Vol. 1. Attachment.* Basic Books, 1969.

C. R. Brewin, M. Cloitre, P. Hyland, M. Shevlin, A. Maercker, R. A. Bryant, A. Humayun, L. M. Jones, A. Kagee, C. Rousseau, D. Somasundaram, Y. Suzuki, S. Wessely, M. van Ommeren, and G. M. Reed. A review of current evidence regarding the ICD-11 proposals for diagnosing PTSD and complex PTSD. *Clinical Psychology Review*, 58:1–15, 2017. doi: 10.1016/j.cpr.2017.09.001.

S. Brown, P. J. Fite, K. Stone, M. J. Richman, and M. Bortolato. Associations between emotional abuse and neglect and dimensions of alexithymia: The moderating role of sex. *Personality and Individual Differences*, 116:176–180, 2017. doi: 10.1016/j.paid.2017.04.049.

J. Cassidy and P. R. Shaver. Contributions of attachment theory and research. In *Handbook of attachment*, pages 3–28. Guilford Press, 3rd edition, 2013.

D. E. Chamberlin. The predictive processing model of emdr. *Frontiers in Psychology*, 10:2267, 2019. doi: 10.3389/fpsyg.2019.02267.

Z. Chen, F. Wang, R. Mu, P. Xu, W. Ruan, and X. Huang. Nrat: Towards adversarial training with inherent label noise. *Machine Learning*, 113:3589–3610, 2024. doi: 10.1007/s10994-023-06437-3.

A. Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013. doi: 10.1017/S0140525X12000477.

M. Cloitre, M. Shevlin, C. R. Brewin, J. I. Bisson, N. P. Roberts, A. Maercker, T. Karatzias, and P. Hyland. ICD-11 complex post-traumatic stress disorder: a new diagnosis. *Journal of Traumatic Stress*, 32(5):699–710, 2019. doi: 10.1002/jts.22446.

P. W. Corrigan, S. B. Morris, P. J. Michaels, J. D. Rafacz, and N. Rüsch. Challenging the public stigma of mental illness: A meta-analysis of outcome studies. *Psychiatric Services*, 63 (10):963–973, 2012. doi: 10.1176/appi.ps.201100529.

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

A. de Jongh, G. N. Groenland, E. t. Broeke, K. E. M. Biesheuvel-Leliefeld, and M. Lehnung. State of the science: Eye movement desensitization and reprocessing therapy. *Journal of Traumatic Stress*, 37(1):8–28, 2024. doi: 10.1002/jts.23002.

C. Dong, L. Liu, and J. Shang. Label noise in adversarial training: A novel perspective to study robust overfitting. In *International Conference on Learning Representations*, 2022.

M. Doucleff. Bringing up a baby can be a tough and lonely job: Alloparents across cultures. NPR Goats and Soda, 2023.

E. H. Emmott and R. Mace. Alloparenting. In *Encyclopedia of Evolutionary Psychological Science*. Springer, 2019. doi: 10.1007/978-3-319-16999-6_2253-1.

V. J. Felitti, R. F. Anda, D. Nordenberg, D. F. Williamson, A. M. Spitz, V. Edwards, M. P. Koss, and J. S. Marks. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults: The adverse childhood experiences (ace) study. *American Journal of Preventive Medicine*, 14(4):245–258, 1998. doi: 10.1016/S0749-3797(98)00017-8.

C. J. Ferguson. Spanking, corporal punishment and negative long-term outcomes: A meta-analytic review of longitudinal studies. *Clinical Psychology Review*, 33(1):196–208, 2013. doi: 10.1016/j.cpr.2012.11.002.

E. T. Gershoff. Corporal punishment by parents and associated child behaviors and experiences: A meta-analytic and theoretical review. *Psychological Bulletin*, 128(4):539–579, 2002. doi: 10.1037/0033-2909.128.4.539.

D. Glaser. Emotional abuse and neglect (psychological maltreatment): A conceptual framework. *Child Abuse & Neglect*, 26(6–7):697–714, 2002. doi: 10.1016/S0145-2134(02)00342-3.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT Press, 2016.

A. Gopnik and H. M. Wellman. Bayesian models of child development. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(2):75–86, 2015. doi: 10.1002/wcs.1330.

C. Hamel, C. Rodrigue, N. Godbout, and M. Hébert. Alexithymia as a mediator of the associations between child maltreatment and internalizing and externalizing behaviors in adolescence. *Scientific Reports*, 14(1):6251, 2024. doi: 10.1038/s41598-024-56909-2.

A. Heilmann, A. Mehay, R. G. Watt, Y. Kelly, J. E. Durrant, J. van Turnhout, and E. T. Gershoff. Physical punishment and child outcomes: A narrative review. *The Lancet*, 398 (10297):355–364, 2021. doi: 10.1016/S0140-6736(21)00582-1.

D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the International Conference on Learning Representations*, 2019. arXiv:1903.12261.

S. B. Hrdy. *Mothers and others: The evolutionary origins of mutual understanding*. Belknap Press of Harvard University Press, 2009.

Q. J. M. Huys, T. V. Maia, and M. J. Frank. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, 19(3):404–413, 2016. doi: 10.1038/nn.4238.

P. Hyland, M. Shevlin, C. R. Brewin, M. Cloitre, A. J. Downes, S. Jumbe, T. Karatzias, J. I. Bisson, and N. P. Roberts. Validation of post-traumatic stress disorder (PTSD) and complex PTSD using the International Trauma Questionnaire. *Acta Psychiatrica Scandinavica*, 136 (3):313–322, 2017. doi: 10.1111/acps.12771.

R. J. Jacoby and J. S. Abramowitz. Inhibitory learning approaches to exposure therapy. *Current Opinion in Psychology*, 2:28–33, 2016. doi: 10.1016/j.copsyc.2014.12.002.

A. P. Kaye, M. G. Rao, A. C. Kwan, K. J. Ressler, and J. H. Krystal. A computational model for learning from repeated traumatic experiences under uncertainty. *Cognitive, Affective, & Behavioral Neuroscience*, 23(3):894–904, 2023. doi: 10.3758/s13415-023-01085-5.

S. Khalfa and C. F. Touzet. Emdr therapy mechanisms explained by the theory of neural cognition. *Journal of Trauma & Stress Disorders & Treatment*, 6(4), 2017. doi: 10.4172/2324-8947.1000173.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and

R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

N. Kriegeskorte and P. K. Douglas. Cognitive computational neuroscience. *Nature Neuroscience*, 21(9):1148–1160, 2018. doi: 10.1038/s41593-018-0210-5.

R. Landin-Romero, A. Moreno-Alcazar, M. Pagani, and B. L. Amann. How does eye movement desensitization and reprocessing therapy work? a systematic review. *Frontiers in Psychology*, 9:1395, 2018. doi: 10.3389/fpsyg.2018.01395.

M. S. Lebowitz and W.-k. Ahn. Effects of biological explanations for mental disorders on clinicians' empathy. *Proceedings of the National Academy of Sciences*, 111(50):17786–17790, 2014. doi: 10.1073/pnas.1414058111.

J. C. G. Lely, G. E. Smid, R. A. Jongedijk, J. W. Knipscheer, and R. J. Kleber. The effectiveness of narrative exposure therapy: A review, meta-analysis and meta-regression analysis. *European Journal of Psychotraumatology*, 10(1):1550344, 2019. doi: 10.1080/20008198.2018.1550344.

A. Lewkowycz, Y. Bahri, E. Dyer, J. Sohl-Dickstein, and G. Gur-Ari. On the training dynamics of deep networks with l2 regularization. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

B. Li, Y. Wang, and W. Liu. Towards understanding catastrophic forgetting in two-layer convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*, 2025.

M. J. MacKenzie, E. Nicklas, J. Brooks-Gunn, and J. Waldfogel. Spanking and children's externalizing behavior across the first decade of life: Evidence for transactional processes. *Journal of Youth and Adolescence*, 44(3):658–669, 2015. doi: 10.1007/s10964-014-0114-y.

J. Marquez, L. Francis-Hew, and N. Humphrey. Protective factors for resilience in adolescence: Analysis of a longitudinal dataset using the residuals approach. *Child and Adolescent Psychiatry and Mental Health*, 17:140, 2023. doi: 10.1186/s13034-023-00687-8.

J. S. Martin, E. J. Ringen, P. Duda, and A. V. Jaeggi. Harsh environments promote alloparental care across human societies. *Proceedings of the Royal Society B*, 287(1933):20200758, 2020. doi: 10.1098/rspb.2020.0758.

A. S. Masten. Ordinary magic: Resilience processes in development. *American Psychologist*, 56 (3):227–238, 2001. doi: 10.1037/0003-066X.56.3.227.

L. R. Miller-Lewis, A. K. Searle, M. G. Sawyer, P. A. Baghurst, and D. Hedley. Resource factors for mental health resilience in early childhood. *Child and Adolescent Mental Health*, 18(1): 44–52, 2013. doi: 10.1111/j.1475-3588.2012.00666.x.

E. R. S. Nijenhuis and O. van der Hart. Dissociation in trauma: A new definition and comparison with previous formulations. *Journal of Trauma & Dissociation*, 12(4):416–445, 2011. doi: 10.1080/15299732.2011.570592.

E. R. S. Nijenhuis, O. van der Hart, and K. Steele. Trauma-related structural dissociation of the personality. *Activitas Nervosa Superior*, 52:1–23, 2010. doi: 10.1007/BF03379560.

Y. Niv and A. Langdon. Reinforcement learning with marr. *Current Opinion in Behavioral Sciences*, 11:67–73, 2016. doi: 10.1016/j.cobeha.2016.04.005.

J. Norman. Alloparenting: A historical perspective on infant "loving" care across cultures. Norland College Repository, 2020.

C. G. Northcutt, L. Jiang, and I. L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021. doi: 10.1613/jair.1.12125.

Q. Pan, S. Chen, and Y. Qu. Corporal punishment and violent behavior spectrum: A meta-analytic review. *Frontiers in Psychology*, 15:1323784, 2024. doi: 10.3389/fpsyg.2024.1323784.

A. Philippsen, S. Tsuji, and Y. Nagai. Simulating developmental and individual differences of drawing behavior in children using a predictive coding model. *Frontiers in Neurorobotics*, 16:856184, 2022. doi: 10.3389/fnbot.2022.856184.

B. Qela, S. Damiani, S. De Santis, F. Groppi, A. Pichiecchio, C. Asteggiano, N. Brondino, A. M. Monteleone, L. Grassi, P. Politi, P. Fusar-Poli, and L. Fusar-Poli. Predictive coding in neuropsychiatric disorders: A systematic transdiagnostic review. *Neuroscience & Biobehavioral Reviews*, 169:106020, 2025. doi: 10.1016/j.neubiorev.2025.106020.

R. J. Quinlan and M. B. Quinlan. Human lactation, pair-bonds, and alloparents: A cross-cultural analysis. *Human Nature*, 19(1):87–102, 2008. doi: 10.1007/s12110-007-9026-9.

F. Raeder, C. J. Merz, J. Margraf, and A. Zlomuzica. The association between fear extinction, the ability to accomplish exposure and exposure therapy outcome in specific phobia. *Scientific Reports*, 10:4288, 2020. doi: 10.1038/s41598-020-61004-3.

H. Rapaport, A. Schettino, P. Sessa, and D. Sauter. Investigating predictive coding in younger and older children. *Developmental Cognitive Neuroscience*, 60:101205, 2023. doi: 10.1016/j.dcn.2023.101205.

M. Rmus, S. D. McDougle, and A. G. E. Collins. Artificial neural networks for model identification and parameter estimation in computational cognitive models. *PLOS Computational Biology*, 20(5):e1012119, 2024. doi: 10.1371/journal.pcbi.1012119.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. doi: 10.1038/323533a0.

M. Schirmer, M. Elsner, B. W. Schuller, and R. D. Findling. The language of trauma: Modeling traumatic event descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, page 773, 2024.

M. E. P. Seligman. *Helplessness: On depression, development, and death.* W. H. Freeman, 1975.

A. D. Shaw, R. L. Sumner, and L. C. S. Berndt. Predictive coding and neurocomputational psychiatry: a mechanistic framework for understanding mental disorders. *Frontiers in Psychiatry*, 16:1713833, 2025. doi: 10.3389/fpsyt.2025.1713833.

M. A. Straus and M. J. Paschall. Corporal punishment by mothers and development of children's cognitive ability: A longitudinal study of two nationally representative age cohorts. *Journal of Aggression, Maltreatment & Trauma*, 18(5):459–483, 2009. doi: 10.1080/10926770903035168.

V. Talwar and K. Lee. A punitive environment fosters children's dishonesty: A natural experiment. *Child Development*, 82(6):1751–1758, 2011. doi: 10.1111/j.1467-8624.2011.01663.x.

C. A. Taylor, J. A. Manganello, S. J. Lee, and J. C. Rice. Mothers' spanking of 3-year-old children and subsequent risk of children's aggressive behavior. *Pediatrics*, 125(5):e1057–e1065, 2010. doi: 10.1542/peds.2009-2678.

Y. Tu, K. Zhou, H. Chen, and M. Gong. Learning with noisy labels via self-supervised adversarial noisy masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6139–6146, 2023. doi: 10.1109/CVPR52729.2023.00594.

M. Ungar. The social ecology of resilience: Addressing contextual and cultural ambiguity of a nascent construct. *American Journal of Orthopsychiatry*, 81(1):1–17, 2011. doi: 10.1111/j.1939-0025.2010.01067.x.

G. M. van de Ven, T. Tuytelaars, and A. S. Tolias. Continual learning and catastrophic forgetting, 2024. arXiv:2403.05175.

O. van der Hart, E. R. S. Nijenhuis, K. Steele, and D. Brown. Trauma-related dissociation: conceptual clarity lost and found. *Australian and New Zealand Journal of Psychiatry*, 38 (11–12):906–914, 2004. doi: 10.1080/j.1440-1614.2004.01480.x.

O. van der Hart, E. R. S. Nijenhuis, and K. Steele. *The Haunted Self: Structural Dissociation and the Treatment of Chronic Traumatization.* W. W. Norton & Company, New York, 2006.

B. A. van der Kolk. *The Body Keeps the Score: Brain, Mind, and Body in the Healing of Trauma.* Viking Press, New York, 2014. ISBN 978-0-670-78593-3.

E. Vanderbilt-Adriance and D. S. Shaw. Protective factors and the development of resilience in the context of neighborhood disadvantage. *Journal of Abnormal Child Psychology*, 36(6): 887–901, 2008. doi: 10.1007/s10802-008-9220-1.

Y. Wang, Z. P. Bian, J. Hou, and L. P. Chau. Convolutional neural networks with dynamic regularization. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2299–2304, 2021. doi: 10.1109/TNNLS.2020.2997044.

D. Yu, Y. Wang, X. Liu, and J. Zhang. Soften to defend: Towards adversarial robustness via self-guided label refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

D. Zhou, T. Liu, B. Han, N. Wang, C. Peng, and X. Gao. Modeling adversarial noise for adversarial training. In *Proceedings of the International Conference on Machine Learning*, 2022.