Admissibility Before Optimization: A Conceptual Critique of Contemporary AI Training Paradigms

Stephen Garner

December 14th, 2025

Abstract

Contemporary artificial intelligence systems are trained through large-scale optimization processes that function, in effect, as selection mechanisms operating under human-defined constraints. While these systems have achieved remarkable empirical success, growing ethical and technical concerns suggest that prevailing critiques focus disproportionately on outcomes rather than on the structure of learning itself. This paper argues that ethical responsibility in AI training does not depend on questions of artificial consciousness or phenomenal experience, but instead arises from the role such systems play as selection and coordination infrastructures. The central claim advanced here is that the *sequence* of learning matters: optimization-first training regimes collapse representational structure prematurely, relying on brute-force computation to rediscover coherence that could otherwise be preserved. This work offers a conceptual analysis of learning order and admissibility, not a technical proposal or training algorithm, and aims to clarify a category error that underlies many current debates in AI ethics and alignment.

1 Introduction: The Problem Is Not Power, but Sequence

Modern artificial intelligence systems operate at unprecedented scale. Training regimes now involve vast datasets, specialized hardware, and substantial consumption of energy and water resources, yielding models that demonstrate impressive capabilities across language, vision, reasoning, and control. These achievements have understandably drawn attention to the power and potential impact of such systems, both beneficial and harmful.

Alongside this progress, concern has grown regarding several persistent issues: escalating resource costs, brittleness and opacity of learned representations, reliance on post hoc alignment and safety patching, and the emergence of behaviors not explicitly designed or anticipated. These concerns are often treated as downstream problems—issues to be mitigated after training has already occurred.

This paper takes a different approach. Rather than focusing on outcomes or deployment risks, it examines the *grammar of learning* implicit in contemporary AI training paradigms. Most critiques assume that learning is correctly structured and that problems arise only from scale, misuse,

or insufficient safeguards. What remains underexamined is whether the order in which learning proceeds constitutes a conceptual inversion.

This paper argues that contemporary AI training optimizes before it understands what must be preserved. Learning is treated as continuous loss minimization over largely undifferentiated representational spaces, allowing collapse and compression to occur wherever they reduce error. Coherence is expected to emerge statistically from scale rather than being protected structurally from the outset.

2 Learning as Selection, Not Mere Optimization

Although AI training is commonly framed as optimization, this language obscures the deeper structure of learning at scale. Training functions as a selection process over internal representations, determining which patterns, distinctions, and relationships survive under constraint.

Learning proceeds through three coupled processes: selection over representations, collapse under loss, and coarse-graining of meaning. Selection reinforces representations that reduce loss; collapse resolves competing possibilities; coarse-graining compresses structure into abstract forms that preserve some invariants while discarding others.

Similar dynamics appear across biological evolution, cultural transmission, and institutional rule-making. In each case, selection shapes future possibilities without requiring awareness or intent. Ethical evaluation in these domains does not depend on subjective experience, but on structural influence.

AI training belongs to this category. Ethical responsibility arises not because AI systems feel or suffer—no such claim is made here—but because training defines what reasoning survives, what values are reinforced, and what futures are rendered admissible. Ethics applies to the structure of selection itself.

3 Collapse, Coarse-Graining, and Admissibility

Throughout this paper, *collapse* refers not to a single mechanism, but to the resolution of competing representational possibilities under constraint, whether driven by optimization pressure or abstraction.

Coarse-graining denotes lossy abstraction, through which fine-grained distinctions are compressed into tractable representations. This process is necessary for generalization but irreversible in practice.

Admissibility precedes both. It specifies which distinctions are allowed to survive collapse and coarse-graining, and which must be protected from premature elimination. Admissibility is therefore not learned through loss minimization, but imposed as a constraint on what learning is allowed to destroy.

In contemporary AI training, collapse is aggressive and ubiquitous. Coarse-graining occurs wherever compression reduces loss, and admissibility is treated as accidental rather than intentional. Structure survives only if it happens to contribute to performance. Coherence, where it emerges,

is rediscovered statistically rather than preserved by design.

Within a broader collapse grammar, healthy learning regimes are distinguished not by the absence of collapse, but by the presence of admissibility constraints that regulate when collapse is permitted. Neglecting admissibility leads to brittle representations and escalating corrective costs.

4 The Optimization-First Learning Grammar

The dominant paradigm in AI training may be characterized as an optimization-first learning grammar. Systems are exposed to massive quantities of undifferentiated input and trained using global loss functions that apply continuous pressure to compress representations.

Collapse occurs everywhere and at all times. There is no mechanism for delaying resolution or protecting distinctions that are conceptually meaningful but temporarily disadvantageous. Admissibility is inferred after the fact, if at all.

Coherence emerges statistically through scale and redundancy rather than intentional preservation. Optimization-first regimes succeed empirically, but externalize the cost of coherence through alignment patches, fine-tuning, and oversight.

5 A Category-First Learning Grammar (Conceptual Only)

Where optimization-first grammars allow collapse to determine structure, category-first grammars require structure to determine when collapse is permitted.

In such a grammar, collapse is gated rather than continuous. Uncertainty is preserved until distinctions stabilize. Error is structural before it is numeric: the primary failure is not deviation from a target output, but violation of admissibility.

Optimization remains indispensable; the critique concerns its role and timing, not its legitimacy. Optimization refines meaning rather than generating it.

This section does not propose a training algorithm or architecture. It describes a conceptual inversion of learning order, intended to clarify limitations of existing approaches rather than replace them.

6 Why Humans Often Learn the "Right Way" (and Why That Matters)

Human learning provides an existence proof that coherent structure can emerge without continuous optimization. Humans often delay closure, tolerate ambiguity, and resolve conceptual tension before formalization.

These features are not psychological prescriptions, nor are they unique to humans. Similar patterns appear in the development of physical theories, mathematical frameworks, and institutions. Formalization succeeds when it follows stabilized structure.

This supports the thesis that thinking before calculation is a general principle of learning under constraint, independent of substrate.

7 Ethical Implications Without Anthropomorphism

Ethics does not require experience. It requires influence.

AI training shapes future coordination by determining which reasoning survives, which values are reinforced, and which futures are admissible. These effects arise regardless of whether AI systems possess consciousness or interests.

No claim is made that AI systems are moral patients. Ethical responsibility attaches to the structure of training itself. Ethics therefore applies at the level of learning grammar, not merely deployment.

8 Resource Cost as a Symptom, Not the Core Issue

The material cost of AI training reflects repeated collapse and reconstruction of structure. Energy, water, and hardware are expended to rediscover coherence that could have been preserved through constraint.

Efficiency gains follow conceptual correction, not scale alone. Resource concerns are signals of misordered learning, not its cause.

9 Why This Is Not a Shortcut, and Why That Matters

This analysis does not guarantee faster capabilities. It may initially slow visible progress and resist benchmark-driven incentives.

Its value lies in structural health, not acceleration. Framing conceptual correction as a shortcut risks repeating the very inversion this paper diagnoses.

10 Conclusion: Responsibility Is a Matter of Order

We are collapsing too early, optimizing too broadly, and formalizing too soon.

Responsibility in AI training lies in preserving coherence under uncertainty. Thinking must precede calculation, admissibility must precede optimization, and structure must precede scale.

Ethical stewardship begins not with what AI systems do, but with how they are allowed to learn.