Constitutional AI Training Lacks Depth-Selective Control

by Nick Clark | Published March 27, 2026 | PDF

Anthropic's constitutional AI represents the most principled approach to alignment training. Explicit constitutional principles guide the model's behavior through training rather than relying solely on example-based RLHF. The approach produces notably well-behaved models. But constitutional training does not govern the depth at which principles are learned. Whether a constitutional principle is absorbed at deep layers that resist fine-tuning or shallow layers that can be easily overridden is an emergent property, not a governed outcome. Training governance provides the depth-selective control that principled training requires.

What Anthropic built

Constitutional AI defines explicit principles that guide model behavior during training. The model is trained to evaluate its own outputs against these principles and revise accordingly. RLHF refines behavior based on human preference data. The combination produces models that are more consistently principled than those trained through RLHF alone. The constitutional approach provides transparency about what principles govern the model's behavior.

The training process applies these principles through the loss function and reward model. The model learns to satisfy constitutional constraints. Which layers of the model absorb which principles, and how deeply those principles are embedded, is determined by the training dynamics rather than governed by the pipeline.

The gap between principled training and depth-governed principles

A constitutional principle learned at shallow layers may be effective during normal operation but vulnerable to fine-tuning attacks or adversarial prompting that accesses deeper representations. The same principle learned at deep layers resists these attacks but may be difficult to update when the principle needs refinement. Depth-selective governance provides structural control: safety-critical principles route to deep, fine-tuning-resistant layers, while adaptable behavioral preferences route to layers that support ongoing refinement.

Provenance tracing becomes particularly valuable for constitutional training. When the model produces an output that appears to violate a constitutional principle, provenance tracing can identify whether the principle was insufficiently learned, whether it conflicts with another learned behavior, or whether the specific input triggered a representation that bypasses the principle's layer.

What training governance enables

With depth-selective gradient routing, constitutional principles are structurally routed to appropriate depth levels. Core safety principles embed at layers that resist modification. Behavioral style preferences embed at adaptable layers. The training pipeline governs not just what principles the model learns but how deeply and how resistant to modification each principle becomes. This gives Anthropic structural control over the robustness hierarchy of its constitutional principles.

The structural requirement

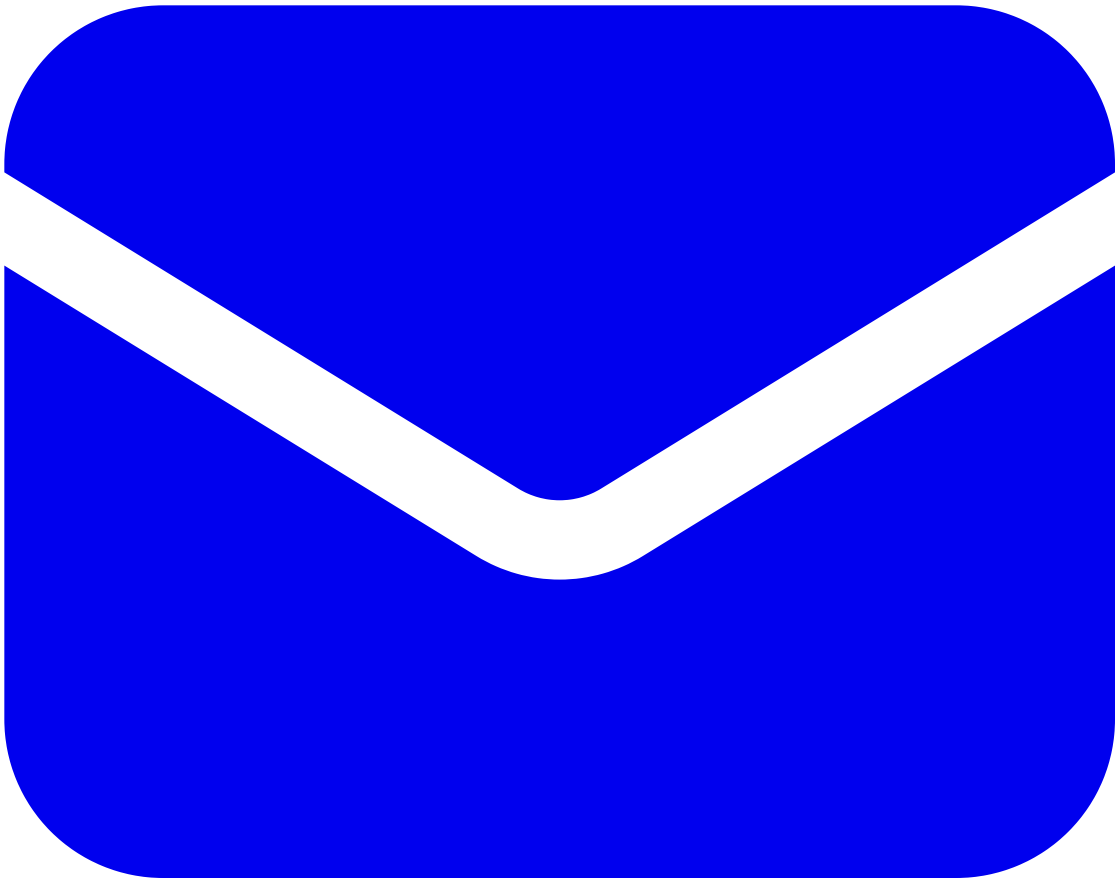
Anthropic's constitutional approach is the most principled training methodology. The structural gap is depth control: governing which layers learn which principles and how resistant each principle is to subsequent modification. Training governance provides the depth-selective routing, entropy-based profiles, and provenance tracing that make constitutional training structurally robust rather than statistically effective.

Training Governance All 21 steps →

Govern what the model learns, at what depth, with what provenance.

Primary Technical Disclosure
○ Depth-Selective Training Governance for Machine Learning Systems
Secondary Technical
○ Training Examples as Proposed Semantic Mutations○ Entropy-Band-Indexed Training Depth Profiles○ Depth-Selective Gradient Routing for Governed Training○ Training-Level Memorization Detection○ Differential Privacy Through Depth-Selective Routing○ Governed Fine-Tuning With Verifiable Provenance○ The Training Loop as a Governed Execution Environment○ Policy-Governed Knowledge Retention and Suppression○ Provenance-Traceable Training Dynamics○ Curriculum-Integrated Depth Scheduling○ Affect-Modulated Training Depth○ Training-Inference Governance Integration○ Training Governance for Human-Relatable Agents
Applications (General)
○ Rights-Compliant Model Training Through Depth-Selective Routing○ Regulated Industry Model Governance With Provenance○ Training Governance for Medical AI○ Training Governance for Legal AI○ Training Governance for Financial Model Training○ Training Governance for Defense AI○ Training Governance for Educational AI Models○ Training Governance for Creative AI
Applications (Specific)
○ OpenAI's Training Pipeline Has No Depth-Selective Governance● Constitutional AI Training Lacks Depth-Selective Control○ Stable Diffusion's Training Has No Provenance Layer○ Midjourney Trains Aesthetics Without Governed Depth○ Scale AI Labels Data Without Governing What Models Learn○ Labelbox Manages Annotation Workflows, Not Learning Dynamics○ Snorkel AI Programs Labels but Does Not Govern Gradient Depth○ Weights & Biases Tracks Experiments, Not Learning Governance○ Determined AI Orchestrates Compute, Not Learning Depth○ MosaicML Optimizes Training Efficiency, Not Learning Governance
Training Governance overview →
AQ
deterministic
autonomy

Last updated: 2026-03-03



-
- nick@qu3ry.net
- 72 28 14 36 01

Invented by Nick Clark | Founding Investors: Devin Wilkie