Constitutional AI Defines Principles Without Cognitive Architecture

by Nick Clark | Published March 27, 2026 | PDF

Anthropic's constitutional AI is the most explicit approach to principled AI behavior. The principles are defined, the model is trained to follow them, and the behavior is evaluated against them. This is more rigorous and transparent than alignment through preference data alone. But constitutional principles are constraints applied to a model. They are not cognitive architecture that embodies those principles through structural dynamics. Human-relatable intelligence provides the architecture where principles emerge from the interaction of cognitive primitives rather than being imposed from outside.

The gap between declared principles and embodied principles

A constitutional principle that the model should be honest is a training objective. A cognitive architecture where integrity is a persistent state variable that tracks behavioral consistency, detects deviation, and triggers self-correction embodies honesty structurally. The first can be trained to approximate honest

behavior. The second cannot be dishonest without its own integrity tracking detecting the deviation. The structural difference is between compliance and constitution in the architectural sense.

Human cognition does not follow principles from a list. Principles emerge from the interaction of emotional state, integrity tracking, empathy, confidence, and planning. A person who is honest is not consulting a principle. Their cognitive architecture produces honest behavior because integrity, empathy, and self-esteem interact in ways that make dishonesty cognitively costly. Human-relatable intelligence reproduces this structural dynamic.

What human-relatable intelligence provides

The ten conditions for human-relatable intelligence define when a computational system is structurally isomorphic to human cognition. The coherence control loop maintains internal consistency. The narrative identity provides continuity. The architectural inversion means governance is not an overlay but the foundation. Anthropic's constitutional principles become parameters within this architecture rather than constraints on a model that lacks it.

The structural requirement

Anthropic's constitutional approach is the most principled AI safety methodology. The structural gap is between principles as training objectives and principles as emergent properties of cognitive architecture. Human-relatable intelligence provides the cognitive dynamics that make principles structural rather than learned, embodied rather than declared. The AI system whose principles emerge from its cognitive architecture is more reliably principled than one trained to follow a list.

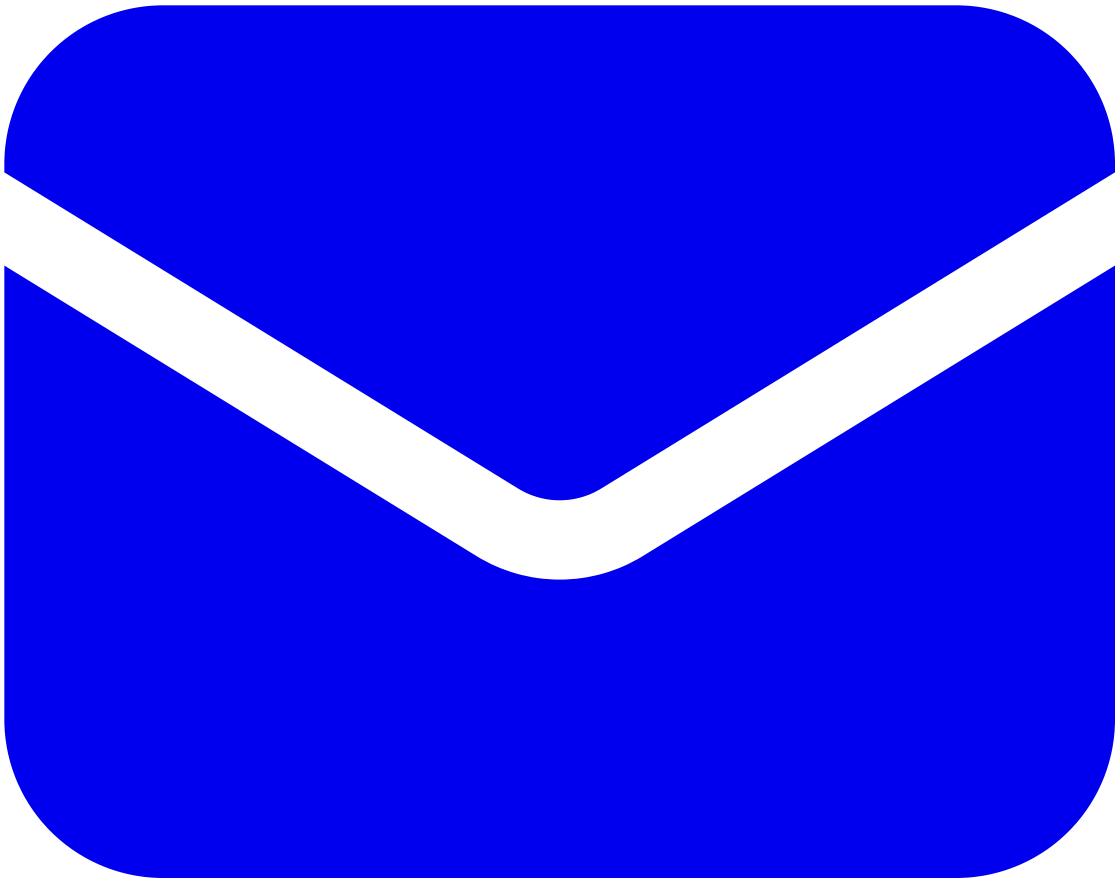
Human-Relatable Intelligence All 21 steps →

The most human-like computer ever built.

Primary Technical Disclosure
○ Human-Relatable Computable Intelligence
Secondary Technical
○ The Cross-Primitive Coherence Engine○ Narrative Identity as Compressed Self-Model○ Ecosystem Participation Credentials From Cognitive History○ Anonymized Governance Telemetry Aggregation○ The Coherence Control Loop: Detection, Recording, Restoration○ The Complete Thirteen-Stage Mutation Lifecycle○ Ten Conditions for Human-Relatable Behavior○ Graceful Degradation With Active-Domain Registry○ Architectural Inversion: Agent Carries State, Substrate Provides Environment○ Sequential Cascade Structures in Cross-Primitive Coherence○ Conformity Attestation: Verifiable Architectural Compliance
Applications (General)
○ Why AI 2.0 Requires Structural Cognition, Not Better Prompts○ The Compliance Case for Cognitive Architecture Under the EU AI Act○ Why Alignment Is Insufficient for Trustworthy AI○ Enterprise Trust Through Architecture, Not Alignment○ Insurance Liability Reduction Through Human-Relatable AI○ Building Consumer Trust in AI Through Cognitive Relatability○ Regulatory Future-Proofing Through Human-Relatable Architecture○ Competitive Differentiation Through Cognitive Architecture
Applications (Specific)
○ OpenAI's Alignment Approach Is Missing Structural Isomorphism● Constitutional AI Defines Principles Without Cognitive Architecture○ DeepMind's Safety Research Lacks Cognitive Isomorphism○ Meta's Open AI Safety Is Missing Cognitive Architecture○ Inflection AI Simulates Empathy Without Structural Coherence○ Adept AI Automates Actions Without Structural Integrity○ Covariant Trains Robot Dexterity Without Cognitive Coherence○ Sanctuary AI Builds Humanoid Form Without Human-Relatable Cognition○ Aleph Alpha Offers Sovereign AI Without Structural Coherence○ Mistral AI Optimizes Efficiency Without Architectural Coherence
Human-Relatable Intelligence overview →
AQ
deterministic
autonomy

Legal

Last updated: 2026-03-03



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

Invented by Nick Clark | Founding Investors: Devin Wilkie