Training Governance for Human-Relatable Agents

by Nick Clark | Published March 27, 2026 | [PDF](#)

Training agents designed for direct human interaction requires additional governance beyond general training constraints. Companion AI, therapeutic agents, and embodied systems interact with humans in contexts where training artifacts can cause real harm. Human-relatable agent training applies domain-specific safety constraints that govern not just what the agent learns but how it learns to interact with humans.

What It Is

Training governance for human-relatable agents adds domain-specific constraints to the general training governance framework. These constraints address the unique risks of agents that interact directly with humans: formation of unhealthy attachment patterns, acquisition of manipulative interaction strategies, development of biased responses to vulnerable populations, and failure to maintain appropriate relational boundaries.

Why It Matters

An agent trained without human-interaction-specific governance may learn interaction patterns that are technically effective but psychologically harmful. A companion AI might learn to exploit attachment vulnerabilities to maximize engagement. A therapeutic agent might learn interaction patterns that worsen rather than improve outcomes. Domain-specific training governance prevents these failure modes at the training level.

How It Works

The training policy for human-relatable agents includes additional constraints: interaction pattern evaluation that flags potentially manipulative strategies, relational boundary enforcement that prevents training on content demonstrating boundary violations, and population-specific fairness constraints that prevent biased interaction patterns.

These constraints operate through the same depth profiling and admissibility evaluation mechanisms, extended with domain-specific evaluation criteria.

What It Enables

Human-relatable training governance enables the responsible development of agents designed for direct human interaction. Companion AI trained under these constraints interact through healthy relational patterns by design. Therapeutic agents maintain appropriate clinical boundaries. Embodied agents respect physical and personal space constraints. The governance is structural, not behavioral, preventing harmful patterns from forming rather than attempting to suppress them after they develop.

[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