Beyond Constitutional AI: A Narrative-Based Methodology for Human-Centered AI Coexistence

Executive Summary

Current AI approaches have important limitations. Constitutional AI struggles when abstract principles like "be helpful and harmless" conflict in real situations. Formal causal inference fails to capture how humans understand cause-and-effect in social contexts. This paper proposes a narrative-based methodology that shapes AI behavior through curated stories, literature, and philosophical works.

Core Insight

Humans rely heavily on narratives alongside direct experience to understand both values and causation. A therapist understands relationship dynamics through case studies; citizens grasp historical events through stories; parents learn that "inconsistent discipline causes behavioral problems" from cultural wisdom. Even without experiencing everything firsthand, we navigate complex moral, social and physical situations through accumulated narratives. This suggests a natural pathway for AI development: teaching systems to produce appropriate behavioral responses by absorbing the moral and causal frameworks embedded in human stories.

The Approach

Following Harari's framework, we treat AI as beneficial alien intelligence – not trying to recreate human consciousness but teaching it to function effectively within human society. The system learns from user-curated character models like Marcus Aurelius (duty, reflection) and Atticus Finch (moral courage) while acknowledging their historical contexts and limitations.

Stories naturally integrate values and causation in ways that abstract rules cannot. "The Tortoise and the Hare" simultaneously teaches moral values (perseverance over arrogance) and causal patterns (consistent effort leads to success). This integration mirrors how humans reason and navigate the world.

Key Advantages

- **Richer than constitutional AI**: Instead of abstract rules, narratives provide concrete examples of navigating moral complexity
- **Human-centered causation**: Captures how people understand cause-and-effect in social and physical systems, not just formal logic
- **Cultural adaptability**: Users can specify different narrative frameworks for different contexts

- **Transparency**: Moral reasoning can be traced to specific narrative sources through attributions like "according to my narrative training"

The Critical Point

Narrative-based value and causality transfer already happens through Language Model training data selection. AI systems currently absorb human stories, biases, and causal assumptions from their training corpus, but this process is opaque and centrally controlled. This methodology makes that process transparent and configurable, allowing different communities to contribute their wisdom traditions while maintaining clear attribution of behavioral influences all in service of AI coexistence as we would hope to achieve with any alien intelligence.

The framework focuses on practical coexistence with current AI systems through behavioral shaping but does not solve for hypothetical superintelligence or consciousness questions. It acknowledges that effective coexistence requires AI to understand both what humans value and how we believe the world works – knowledge naturally embedded in the stories we tell.