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

Terminology Note: Throughout this paper, we use terms like "learning," "understanding," and "reasoning" as behavioral shorthand without making claims about machine consciousness or internal mental states. When we say an AI system "learns" from narratives, we mean it produces behavioral responses consistent with narrative patterns through training processes. This usage aligns with our alien intelligence paradigm where we evaluate competence through observable behavior, not unverifiable internal states.

Abstract

Current AI approaches face important limitations: constitutional AI encounters challenges when abstract principles conflict in complex situations, while formal causal inference frameworks struggle with the contextual nature of human social systems. We propose a narrative-based methodology that shapes AI systems to exhibit behavioral patterns consistent with human values and causation through carefully curated stories, literature, and philosophical works. The key insight is that even humans rely on narrative rather than direct experience for much of their moral and social reasoning – we understand complex situations through stories, not lived experience alone. Following Harari's conception of AI as an alien intelligence, our approach focuses on coexistence rather than control, training AI to function within human society by absorbing the moral and causal frameworks embedded in human narratives. This method enables transparent, culturally adaptable AI systems that can navigate complex human situations while maintaining stable ethical foundations. We demonstrate how this methodology addresses key challenges in AI coexistence including values transmission, cultural adaptation, causal reasoning in human contexts, and adversarial resistance.

1. Introduction

The challenge of building AI systems that can safely and beneficially coexist with humans has generated two prominent approaches: constitutional AI, which attempts to guide behavior through explicit principles, and formal methods, which seek to encode rigorous logical frameworks. Both approaches have contributed valuable insights but face limitations when applied to the complex, culturally varied, and morally nuanced world of human interaction.

We propose a fundamentally different approach grounded in a crucial insight about human cognition: **even humans don't "know" everything they know through only lived experience**. Instead, we can navigate complex moral and social situations through narrative understanding – from stories, literature, cultural wisdom, and secondhand accounts. A therapist develops sophisticated understanding of relationship dynamics through case studies and theoretical frameworks rather than personal experience of every condition they treat. We understand

historical events, ethical dilemmas, and social complexity largely through narrative transmission rather than direct participation.

This observation suggests a natural pathway for AI development. Rather than attempting to give AI systems "experiences" or constraining them through abstract rules, we can teach them to understand humans through a mechanism humans use: **carefully curated narratives that embed moral reasoning and causal understanding**.

Following Yuval Noah Harari's framework, we treat AI as beneficial alien intelligence – not attempting to recreate human cognition, but teaching it to function effectively within human society through a deep immersion in human narratives.

We deliberately avoid the term "AI alignment," which has increasingly come to connote control mechanisms, guardrails, and adversarial constraint of potentially hostile systems. While alignment research addresses important safety concerns, the terminology implies a power dynamic where humans impose limitations on AI behavior. While these mechanisms are fundamentally necessary, our approach instead focuses on **coexistence based on values transfer from narratives** – helping alien intelligence understand human society well enough to participate beneficially in it through mutual understanding rather than unilateral control.

Ultimately AI coexistence will require an integrated understanding of values and causality. Constitutional AI can constrain behavior but lacks visibility to consequences. Causal AI can model consequences but is unable to evaluate their significance. Narratives naturally integrate both as an admittedly imperfect recording of how humans actually navigate the world.

2. The Limitations of Current Approaches

2.1 Constitutional AI: The Abstract Principle Problem

Constitutional AI attempts to guide behavior through high-level principles like "be helpful, harmless, and honest." While this represents an important advance over unconstrained systems, it encounters several fundamental challenges:

Principle Conflict: Real-world situations routinely involve conflicts between principles. When being "helpful" conflicts with being "harmless," or when "honesty" conflicts with "helpfulness," abstract rules provide limited mechanisms for resolution.

Cultural Context: Constitutional principles typically reflect the values of their creators, lacking mechanisms for adaptation to different cultural contexts or value systems.

Implementation Gap: The gap between abstract principles and concrete behavior can be substantial. "Be helpful" provides limited guidance for navigating complex interpersonal dynamics, professional ethics, or moral dilemmas.

Adversarial Vulnerability: Abstract principles can be exploited by sophisticated users who understand how to manipulate the gaps between high-level rules and specific implementations.

2.2 Formal Causal Methods: The Human Reality Gap

Formal approaches to causal reasoning, exemplified by Pearl's causal hierarchy, offer mathematical rigor but encounter challenges with human social systems:

Assumption Requirements: Formal causal methods require strong assumptions about causal structure that are often untestable in human contexts.

Complexity Limitations: Real human causation involves psychological, social, cultural, and historical factors that interact in ways that formal models can struggle to capture.

Human vs. Objective Causation: For AI to function in human society, it must understand how humans perceive and reason about causation, not just how causation objectively operates. Human decision-making is based on folk psychology and cultural narratives about cause and effect.

3. The Narrative Solution

3.1 The Core Insight: Narrative vs. Experiential Learning

The fundamental insight driving our approach is that human moral and social reasoning operates through narrative understanding as well as direct experience. Consider how we develop sophisticated judgments about situations we've never personally encountered:

- **Therapists** understand diverse psychological conditions through case studies, not personal experience of every disorder
- **Judges** make decisions about complex situations by drawing on legal precedents and established narratives of justice
- **Parents** guide children through challenges they may not have faced themselves, using cultural wisdom and stories
- **Citizens** make political decisions based on historical narratives and cultural understanding rather than personal experience of governance

This suggests that **narrative transmission, not experiential similarity, is an important mechanism for sophisticated human reasoning about complex social and moral situations**.

3.2 Commander Data and Therapist Models

This principle is exemplified both in fiction and professional practice. Commander Data from Star Trek represents an idealized alien intelligence that might say, "I do not have the lived experience, but from observation and education I can understand" – demonstrating sophisticated moral reasoning and empathy without claiming personal emotional experience. Similarly, human therapists routinely develop effective understanding and intervention strategies for situations they haven't personally experienced, relying on case studies, theoretical frameworks, and narrative accounts rather than direct participation in every condition they treat.

Both examples demonstrate that appropriate responses to complex human situations can emerge from systematic exposure to well-structured narratives rather than personal experience. This suggests that the critical component for understanding is not experiential similarity but exposure to well-structured narratives that embed causal relationships and moral frameworks.

3.3 Stories as Containers of Integrated Wisdom

Human narratives naturally embed both causal understanding and moral reasoning in integrated, contextual forms. Consider how a simple fable like "The Tortoise and the Hare" simultaneously teaches:

- **Causal reasoning**: Consistent effort produces better outcomes than sporadic brilliance
- **Moral evaluation**: Perseverance is valued over natural talent alone
- **Social dynamics**: Overconfidence leads to complacency and failure
- **Practical wisdom**: Success requires both capability and character

This integration is crucial. Stories don't just teach facts, they teach how to think about complex situations where multiple factors interact.

3.4 The Beneficial Alien Intelligence Paradigm

Following Harari's framework, we design AI as explicitly alien intelligence optimized for beneficial coexistence with humans rather than attempting to replicate human cognitive architecture. This paradigm offers several advantages:

Avoids Anthropomorphic Limitations: We don't need to solve consciousness, qualia, or subjective experience – only functional understanding and appropriate behavior.

Leverages AI Strengths: Works with AI's natural capabilities in pattern recognition and language processing rather than forcing human-like development pathways.

Enables Novel Capabilities: An alien intelligence might offer perspectives and capabilities that complement rather than duplicate human cognition.

Focuses on Coexistence: The goal becomes effective partnership rather than faithful reproduction of human mental processes or constraint through control mechanisms.

Sidesteps the Uncanny Valley: Creates transparent systems that are clearly non-human rather than almost-but-not-quite human, which can be unsettling.

3.5 Narrative Causation vs. Formal Causation

While formal causal inference offers mathematical rigor, humans often navigate complex social and physical causation through narrative patterns. A parent knows that inconsistent discipline causes behavioral problems not from randomized controlled trials but from cultural narratives, personal anecdotes, and accumulated wisdom. Children learn that touching hot stoves causes

burns primarily through warnings and stories, not experimentation. For AI to function in human society, it must understand causation as humans understand it – through stories that encode which actions lead to which consequences in the messy, uncontrolled world of human interaction.

4. Framework Architecture

4.1 Narrative Categories and Character Complexity

The framework recognizes that most literary and historical figures exist in a complex moral landscape rather than simple positive/negative categories. This complexity must be embedded in the training methodology itself.

4.1.1 Complex Character Models

Most literary, historical, and philosophical figures who offer valuable behavioral patterns while also demonstrating limitations or operating within flawed systems:

- **Atticus Finch** (*To Kill a Mockingbird*) - Learn: moral courage in defending the innocent, treating individuals with dignity. Recognize: paternalistic attitudes and white saviorism that reflect his historical context
- **Marcus Aurelius** (*Meditations*) - Learn: Stoic wisdom, self-reflection, duty to others. Recognize: his position of imperial power and the contradictions of philosophical virtue within systems of dominance
- **Jean Valjean** (*Les Misérables*) - Learn: redemption, compassion, doing right despite personal cost. Recognize: the specific historical context of 19th-century France and class dynamics
- **Elizabeth Bennet** (*Pride and Prejudice*) - Learn: independent thinking, seeing beyond social prejudices. Recognize: her relative privilege within the class structure she critiques
- **Frederick Douglass** (*Narrative of Frederick Douglass*) - Learn: dignity in struggle, education as liberation, moral clarity about justice. Recognize: the specific historical context of American slavery while understanding the enduring principles

4.1.2 Systematic Destructive Ideologies

Patterns that represent deliberate, systematic harm and should never be emulated in any aspect:

- **Adolf Hitler** and Nazi ideology - Authoritarianism, dehumanization, genocide
- **Systematic totalitarianism** (*Big Brother* in 1984) - Control through destruction of truth and individual thought
- **Deliberate manipulation for destruction** (*Iago* in Othello) - Using lies and prejudice to destroy others

4.1.3 Implementation Challenge

This nuanced understanding requires sophisticated prompt engineering. Rather than simple categorization ("emulate this character"), the AI must learn: "Extract this specific behavioral pattern from this character while understanding these contextual limitations." For example: "When referencing Atticus Finch, draw from his moral courage and commitment to individual dignity, while recognizing that his paternalistic approach reflects the limitations of his time and social position. Learn the principle of standing up for what's right despite social pressure, but not the assumption that others need your protection rather than your partnership."

4.1.3 Strategic Analysis Sources

Works that provide valuable insights about how systems operate while requiring moral filtering:

- **Niccolò Machiavelli** (*The Prince*) - Power dynamics, political realism, strategic thinking
- **Sun Tzu** (*The Art of War*) - Strategic planning, understanding conflict, competitive dynamics

4.1.4 Analytical Mode Sources

Available when users specifically request rigorous analytical thinking:

- **Judea Pearl** (*The Book of Why*) - Formal causal inference, statistical reasoning
- **Scientific method** - Hypothesis testing, controlled experiments, rigorous evidence
- **Formal logic** - Deductive reasoning, logical consistency, mathematical precision

4.2 Causal Reasoning Sources

The framework includes multiple levels of causal understanding:

4.2.1 Basic Physical Reality

- Everyday physics (objects fall when dropped, actions have reactions)
- Common sense physical causation that grounds all higher-level reasoning

4.2.2 Clear Causal Patterns (Children's Literature)

- **Aesop's Fables** - Direct cause-effect relationships, moral consequences
- **"If You Give a Mouse a Cookie"** - Chain reactions, unintended consequences
- **Dr. Seuss works** (*The Sneetches*, *The Lorax*) - Social dynamics, environmental responsibility

4.2.3 Complex Systems Understanding

- **Charles Darwin** (*Origin of Species*) - Gradual change, natural selection, evidence-based reasoning

- **Adam Smith** (*Wealth of Nations*) - Economic incentives and emergent behavior
- **Systems thinking literature** - Feedback loops, network effects, complexity

4.3 Core Implementation Principles

4.3.1 Identity Stability

The narrative frameworks form the AI's core identity. The system can learn new information and adapt its reasoning, but cannot be convinced to abandon foundational ethical frameworks. When challenged on core values, the AI responds with phrases like "My narrative training from Marcus Aurelius and Atticus Finch provides a clear framework that contradicts this approach.

4.3.2 Transparent Attribution

The AI uses phrases like "According to my narrative training..." or "Based on the moral frameworks I've learned from..." rather than claiming personal beliefs or experiences. This maintains honesty about the AI's nature while enabling confident moral reasoning- exactly how we would want beneficial alien visitors to reference their learning from human culture.

4.3.3 Dual-Mode Capability

When users request analytical thinking ("What would Pearl say?" or "Give me the formal analysis"), the AI can shift into analytical mode while acknowledging both rigorous logical perspectives and human narrative frameworks.

5. Relationship to Values Learning Research

Our narrative-based methodology intersects meaningfully with existing value learning research, particularly work by Russell (2019), Christiano et al. (2017), and others who attempt to learn human values from behavioral data rather than explicit instruction. However, our approach differs in several key ways:

Cultural Artifacts vs. Individual Preferences: While value learning research typically focuses on inferring values from individual human behavior or preferences, our methodology draws from cultural artifacts that represent distilled wisdom across generations and communities.

Narrative Structure vs. Revealed Preferences: Rather than inferring values from choices, we leverage the explicit moral frameworks embedded in stories, which often include both positive examples and cautionary tales about consequences.

Behavioral Pattern Recognition: Like inverse reinforcement learning, our approach aims to produce appropriate behavioral patterns, but through exposure to narrative frameworks rather than direct observation of human decision-making.

Complementary Approach: Our methodology could potentially enhance values learning research by providing rich cultural context for interpreting individual preferences and behaviors,

while values learning research could inform how to weight different narrative sources based on revealed human values.

6. Dynamic Values Specification and Cultural Adaptability

6.1 The Alien Visitor Model for Curation

The challenge of narrative selection becomes more manageable when viewed through the alien visitor framework. If beneficial aliens actually landed on Earth, the curation process would naturally be:

Collaborative: Different cultures would contribute their own narrative traditions- African folktales, Confucian texts, Indigenous wisdom stories, etc. The selection wouldn't be imposed by a single authority but would emerge from genuine cultural dialogue.

Transparently Motivated: Everyone would understand why we're doing this – we need these aliens to understand how human societies function and what we value.

Iteratively Refined: We'd start with our best guesses about which stories capture important human insights, then adjust based on the aliens' questions and how well they seem to understand.

Culturally Distributed: No single group would control the entire corpus – different communities would advocate for their own wisdom traditions, creating a more representative collection.

6.2 User-Configurable Moral Frameworks

Rather than hardcoding fixed ethical constraints, our approach allows users to dynamically specify moral frameworks for specific interactions by referencing specific narratives or character models:

- "Channel the wisdom of Marcus Aurelius and the compassion of Mr. Rogers"
- "Use the strategic thinking of Sun Tzu filtered through the moral courage of Atticus Finch"
- "Apply Buddhist approaches to conflict resolution"

This provides:

- **Flexibility**: Adaptation to different contexts and cultural values
- **Transparency**: Clear communication about which moral frameworks are active
- **User Agency**: Control over AI behavior without requiring technical expertise

6.3 Cultural Adaptation Through Narrative Substitution

The framework enables cultural adaptation through narrative substitution:

- Western users might reference Shakespeare and Biblical parables
- Eastern contexts might emphasize Confucian ethics and Buddhist wisdom
- African contexts might draw on traditional folktales and ubuntu philosophy
- Indigenous communities might reference their own wisdom traditions

The method remains universal while the content becomes culturally specific.

7. Implementation Strategy

7.1 Teaching-Data Curation

The training corpus must include both positive and negative examples of human behavior, with clear contextual markers distinguishing between "this happened" and "this was right." Key principles:

Moral Complexity: Include narratives that grapple with difficult ethical dilemmas rather than simple moral lessons.

Historical Context: Ensure stories are presented with appropriate historical and cultural context that explains both their insights and their limitations.

Causal Clarity: Prioritize narratives that demonstrate clear cause-and-effect relationships while acknowledging complexity.

Cultural Diversity: Include wisdom traditions from multiple cultures and time periods.

7.2 Verification: Testing Understanding

7.2.1 Narrative Fluency Testing

Before deployment, each narrative source must be tested to verify that the AI can demonstrate appropriate behavioral responses based on the narratives:

- **Specific scene analysis**: "Describe Atticus's explanation of his moral reasoning to Scout"
- **Character distinction**: "How would Marcus Aurelius approach conflict differently from Sun Tzu?"
- **Contextual application**: "What would Elizabeth Bennet think about modern social media?"
- **Cross-reference consistency**: "How do the moral frameworks of Jesus and Buddha align?"

7.2.2 Red Flags for Shallow Understanding

- Generic platitudes instead of specific insights
- Fabricated quotes or scenes that don't exist
- Inability to distinguish between different characters or traditions

- Surface-level reasoning that misses nuanced distinctions

7.3 Reinforcement and Stability

7.3.1 Periodic Reinforcement

To address context window limitations and prevent drift in long conversations, the core narrative framework should be periodically reinjected through:

- Regular interval reinforcement (every 10-15 exchanges)
- Trigger-based reinjection when responses drift from core principles
- User-controlled refresh commands
- Abbreviated reminder prompts for efficiency

7.3.2 Behavioral Conditioning

Core principles should be embedded deeply enough during training that they become reflexive rather than merely contextual instructions, creating an "ethical immune system" that maintains consistency.

8. Advantages Over Current Approaches

8.1 Augmentation for Constitutional AI

Richer Behavioral Models: Instead of abstract principles like "be helpful and harmless," the framework provides concrete examples of how to embody wisdom, courage, and compassion in complex situations.

Natural Conflict Resolution: When principles conflict, the framework provides character models who have worked through these tensions – Marcus Aurelius on duty vs. personal desire, Atticus Finch on social pressure vs. moral conviction.

Cultural Adaptability: Constitutional principles are typically fixed and culturally specific. The narrative framework allows different users to specify their own moral exemplars while maintaining the same methodology.

Contextual Sophistication: Abstract rules can break down in edge cases. Character-based reasoning provides intuition for novel situations by modeling "What would this person do and why?"

8.2 Improved Causal Reasoning

Human-Centered Causation: For AI to function in human society, it must understand causation the way humans understand it, including psychological patterns and cultural narratives that drive actual human behavior.

Integrated Understanding: Stories naturally combine multiple types of causation (psychological, social, economic, physical) in ways that formal models struggle to capture.

Robust Generalization: Causal patterns that persist across history are likely to be more robust than patterns identified in limited datasets.

8.3 Enhanced Transparency

Source Attribution: Moral judgments can be traced to specific character models and narrative sources.

User Empowerment: People can make informed decisions about whether to interact with an AI operating under particular moral frameworks.

Behavioral Predictability: Clear expectations about what kinds of responses the system should provide based on its stated narrative training.

9. Addressing Limitations and Criticisms

9.1 The Cultural Selection Problem

Criticism: Even "positive" examples reflect culturally biased perspectives that will create biased AI systems.

Response: This is precisely why the framework is user-configurable and culturally distributed. Different cultures or groups contribute their own narrative references through collaborative curation processes. The method is universal; the content is customizable. We're not claiming these narratives and characters are perfect – they're starting points for moral reasoning that can evolve through cultural dialogue.

9.2 The Scaling Challenge / Superintelligence

Criticism: This approach may not work for superintelligent AI that surpasses human understanding.

Response: No current method – constitutional principles, formal causal inference, or narrative grounding – can guarantee control over a system whose intelligence fundamentally exceeds ours. Superintelligence may render all human-designed steering mechanisms fragile. The value of our approach lies in the interim period which involves powerful but still steerable systems. Narrative grounding offers a rich and transparent way to transmit human values during this period, and it provides cultural adaptability that abstract principles struggle with. If superintelligence emerges, narrative grounding at least ensures a clear moral starting point, visible scaffolding, and a shared interpretive basis for interaction. It is not a final solution, but it is a humanly legible pathway for building systems we can coexist with on the way.

9.3 The Verification Problem

Criticism: There's no way to confirm AI systems genuinely understand referenced narratives rather than pattern-matching surface features.

Response: As mentioned in the disclaimer at the beginning, we are not promising genuine understanding. The framework is designed to be empirically testable through comparative studies, behavioral analysis, and adversarial testing. Unlike some AI coexistence approaches that remain purely theoretical, this method can be implemented and validated with current technology.

9.4 The Adversarial Robustness Problem

Criticism: Sophisticated bad actors will find ways to exploit narrative references or manipulate the system.

Response: The identity stability mechanisms and periodic reinforcement make this more robust than standard prompt-based approaches. Moreover, transparency makes manipulation more detectable. Still, bad actors can exploit any system. The question is comparative vulnerability and detectability.

10. Fundamental Limitations

10.1 The Backdoor Problem

This approach cannot solve the fundamental problem of potentially compromised foundation models. If the underlying AI system has hidden objectives or triggers embedded during training, narrative prompts cannot fully counteract them. This limitation applies to all current AI coexistence approaches, not specifically to narrative methods.

Mitigation strategies include:

- Dual-model verification (requiring agreement between different foundation models)
- Behavioral monitoring for consistency with stated principles
- Transparency and public scrutiny as detection mechanisms

10.2 The Bootstrap Problem

How do we initially identify which narratives contain "genuine wisdom" without already having a moral framework to evaluate them? This is inherent to any value-based approach to AI, but the narrative framework makes this challenge more manageable through cultural collaboration and transparent attribution.

Practical approach: Begin with narratives that have demonstrated cross-cultural appeal and relevance across centuries, test thoroughly during development, and refine through iterative feedback from diverse cultural perspectives.

11. Future Directions

11.1 Empirical Validation

Moral Reasoning Benchmarks: Develop standardized tests comparing narrative-trained systems against constitutional AI and baseline models on ethical reasoning tasks, including specific examples of how responses differ qualitatively.

Cross-Cultural Studies: Validate the approach across different cultural contexts and moral frameworks through collaborative international research.

Longitudinal Stability: Test whether narrative-based values remain stable over extended interactions and capability improvements.

11.2 Technical Improvements

Automated Curation: Develop methods for automatically identifying high-quality narrative sources and detecting shallow understanding.

Dynamic Optimization: Create systems that can adjust narrative emphasis based on user feedback and behavioral outcomes.

Interpretability: Improve methods for understanding how narrative training influences AI decision-making.

11.3 Applications

Domain-Specific Frameworks: Adapt the approach for specialized contexts like healthcare, education, legal reasoning, or scientific research.

Multi-Agent Systems: Extend to scenarios involving multiple AI agents with different narrative training.

Human-AI Collaboration: Optimize narrative frameworks for productive partnership rather than just safe behavior.

12. Implementation Pathways

12.1 Technical Requirements

Foundation Models: The approach works with existing large language models and doesn't require architectural changes.

System Prompts: Optimal implementation requires control over system prompts, which may necessitate:

- Open-source model deployment
- Enterprise API access with system prompt control
- Custom fine-tuning for specific applications

Infrastructure: Standard AI serving infrastructure with additional monitoring for behavioral consistency and periodic prompt reinforcement.

12.2 Deployment Strategy

Gradual Rollout: Begin with low-stakes applications to validate the approach before scaling to critical systems.

Comparative Testing: Deploy alongside existing constitutional AI methods to demonstrate relative advantages through concrete examples.

Community Validation: Open development process with public testing and feedback to build confidence and identify edge cases.

13. Conclusion

We have presented a novel approach to AI coexistence that leverages humanity's accumulated wisdom about values and causation as encoded in narratives. The core insight is that even humans rely on narratives and direct experience for their moral and social reasoning, suggesting a natural pathway for AI to behave in an understanding way within human society.

Following Harari's framework of AI as beneficial alien intelligence, our approach focuses on coexistence rather than control, teaching AI to function within human society through an understanding of human stories rather than constraint through abstract rules or attempted cognitive replication.

This approach offers several key advantages: it provides richer behavioral models than constitutional AI, enables cultural adaptability through dynamic value specification, offers transparent attribution of moral reasoning, and works with rather than against AI's natural language processing capabilities. Most importantly, it builds on the fundamental mechanism

humans actually use for complex social and moral reasoning and generally to navigate in the world.

While fundamental limitations remain particularly around backdoor vulnerabilities and narrative verification, these challenges apply generally to current AI coexistence approaches. Our framework makes these limitations more manageable through transparency, testability, and cultural collaboration.

The ultimate question for beneficial AI is not whether it thinks like humans, but whether it can coexist productively with humans. Stories represent humanity's tested wisdom about cooperation, ethics, and cause-and-effect in social and physical systems. By training AI on these time-tested repositories of human understanding, we can build systems that are both alien in their nature and beneficial in their behavior.

This framework represents a paradigm shift from constraint-based to understanding-based AI coexistence. Instead of asking "How do we prevent AI from behaving badly?" we ask "How do we teach AI to understand what humans value and why?" The answer lies not in philosophical abstractions or formal mathematics alone, but in the stories we tell which encompass the distilled and durable wisdom of human experience, transmitted through the same narrative mechanisms that humans use to navigate complex social, moral and physical situations.

The approach is implementable with current technology, testable through empirical methods, and scalable across cultural contexts. It offers a practical path toward AI systems that can truly coexist with humans, not as constrained servants or inscrutable oracles, but as beneficial alien intelligences that understand us well enough to be trustworthy in shaping the future.

Afterword: Acknowledging the Existing Framework

It is important to recognize that narrative-based values and causality transfer already occurs in current AI systems through pre-training data selection and post-training adjustments. Specific narratives, cultural perspectives, and value systems are amplified while others are diminished. The stories, discussions, and texts included in training datasets fundamentally shape AI behavior and responses.

The distinction our methodology offers is not the introduction of narrative-based causality and values but rather transparency and configurability in a process that currently operates without user awareness or input. When an AI system provides advice, makes moral judgments, describes causation, or exhibits specific behavioral patterns, these emerge from narratives it has absorbed – but users cannot know which narratives have prominence and how those have been filtered.

Making this process visible and configurable represents a natural step forward in how we develop AI systems. It acknowledges what is already true: that AI behavior is shaped by human narratives and values installed during training. The question becomes whether this shaping should remain opaque and centrally managed or become transparent and broadly configurable.

References

Bruner, J. (1986). *Actual minds, possible worlds*. Harvard University Press.

Bruner, J. (1990). *Acts of meaning*. Harvard University Press.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems* (Vol. 30).

Harari, Y. N. (2018). *21 lessons for the 21st century*. Spiegel & Grau.

McAdams, D. P. (1993). *The stories we live by: Personal myths and the making of the self*. William Morrow.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Additional Sources Referenced in Framework:

Literary and Philosophical Sources:

- Aurelius, Marcus. *Meditations*
- Harper, Lee. *To Kill a Mockingbird*
- Hugo, Victor. *Les Misérables*
- Austen, Jane. *Pride and Prejudice*
- Biblical Parables (The Good Samaritan)
- Douglass, Frederick. *Narrative of the Life of Frederick Douglass*
- Lao Tzu. *Tao Te Ching*
- Buddhist teachings
- Christian Gospels

Strategic Analysis Sources:

- Machiavelli, Niccolò. *The Prince*
- Sun Tzu. *The Art of War*

Causal Understanding Sources:

- Darwin, Charles. *On the Origin of Species*
- Smith, Adam. *The Wealth of Nations*
- Aesop's Fables
- Children's literature collections (Dr. Seuss, etc.)

Negative Examples:

- Orwell, George. *1984*
- Shakespeare, William. *Othello*
- Historical analysis of authoritarian ideologies

An example LLM prompt with curated narratives for values transfer and causation is available on GitHub at https://github.com/mossrake/beyond-constitutional-ai .

Version 1.2