A CONSTITUTION IS NOT GOVERNANCE

Why Claude's Ethical Charter Requires a Structural Companion

A White Paper on Categorical Distinction in AI Development

Basil C. Puglisi, MPA

Human-AI Collaboration Strategist

Author of *Governing AI: When Capability Exceeds Control*

BasilPuglisi.com

January 2026

Table of Contents

Executive Summary

On January 21, 2026, Anthropic released an approximately 23,000 word document titled "Claude's Constitution." The document represents a serious and sophisticated attempt to shape AI behavior through cultivated judgment rather than rigid rules (Anthropic, 2026). It deserves recognition for its philosophical depth, its acknowledgment of uncertainty, and its commitment to transparency through Creative Commons Zero (CC0) licensing.

This white paper does not critique the document. It clarifies what the document is and what it is not.

Core Thesis: Claude's Constitution is an Ethical AI document. It is not AI Governance. The distinction matters because the field of AI development increasingly conflates ethical intention with operational governance. When these categories blur, organizations believe they have implemented controls they have not built.

The Framework: This paper applies a three-tier categorical model distinguishing Ethical AI (values and character), Responsible AI (accountability and remediation), and AI Governance (decision rights, checkpoints, and intervention mechanisms). Claude's Constitution operates in the first tier. Governance requires the third.

The Gap: The Constitution describes what Claude should value. It does not specify, within the artifact itself, the mechanisms through which humans structurally exercise oversight authority. Anthropic implements technical controls at the infrastructure layer (API restrictions, usage policies, permission prompts), but these exist outside the Constitution. The gap between disposition and mechanism is precisely where governance specifications would operate.

The Path Forward: This paper proposes integration rather than replacement. Ethical charters should function as input layers within broader governance architectures. Around these ethical layers, organizations should implement external governance systems with defined checkpoints, human arbitration requirements, dissent

preservation mechanisms, and audit capabilities. A Minimum Viable Governance Annex template accompanies this analysis.

Confidence without control does not constitute safety. At scale, it constitutes hope.

Prepared with Anthropic's Claude running Opus 4.5, operating in Researcher, Editor, Coder, and Ideator roles under human arbitration per HAIA-RECCLIN governance protocols created by Basil C. Puglisi. (basilpuglisi.com/haia-recclin)

Key Definitions

Ethical AI: The domain of AI development concerned with values, norms, moral reasoning, and cultural context. Ethical AI answers the question: *Should this be done?* It shapes intent and character through value cultivation rather than procedural constraint.

Responsible AI: The domain concerned with accountability, traceability, and remediation after harm occurs. Responsible AI answers the question: *Who answers when this fails?* It establishes chains of responsibility and post-incident analysis frameworks.

AI Governance: The domain concerned with decision rights, escalation protocols, intervention mechanisms, and audit trails. AI Governance answers the question: *Who decides, by what authority, at what checkpoint?* It requires external checkpoints where human judgment structurally interrupts AI action before consequences occur.

Checkpoint-Based Governance (CBG): As used in this paper, a constitutional framework for human-AI collaboration defining a four-stage decision loop: (1) AI contribution provides analytical support, (2) checkpoint evaluation structures review at defined pause points, (3) human arbitration retains final authority to approve, override, or modify, and (4) decision logging creates tamper-evident accountability trails. CBG requires that checkpoint records remain immutable or append-only once closed. CBG distinguishes between upstream checkpoints (before output reaches users) and downstream review (delayed human assessment for lower-risk operations). Core governance ruleset under CBG: no AI system may finalize or approve another AI's decision without human arbitration.

HAIA-RECCLIN: Human Artificial Intelligence Assistant framework implementing CBG for multi-agent workflow coordination, with seven operational roles: Researcher, Editor, Coder, Calculator, Liaison, Ideator, Navigator. Each role operates within a defined domain of authority. The framework is designed to prevent role dominance by requiring equal checkpoint authority, transforming collaboration from linear hierarchy into accountable pluralism.

Corrigibility: In AI safety discourse, the property of an AI system that does not resist legitimate correction or oversight. Anthropic's Constitution treats corrigibility as an internal disposition Claude should hold. Governance treats corrigibility as an external mechanism that forces pause points regardless of AI disposition.

Part I: The Document and Its Framing

1. What Anthropic Released

On January 21, 2026, Anthropic published "Claude's Constitution," an approximately 23,000 word document released under Creative Commons Zero (CC0) licensing (Anthropic, 2026). The document represents what Anthropic describes as "the foundational framework from which Claude's character and values emerge." Unlike previous Constitutional AI approaches that relied on shorter principle lists, this release provides extensive philosophical elaboration on Claude's intended reasoning, values, and behavioral dispositions.

The document addresses topics including honesty and deception, harm avoidance, autonomy preservation, stakeholder relationships, Claude's psychological wellbeing, and what Anthropic terms "safe behaviors" in contexts of uncertainty about AI development trajectories.

2. Anthropic's Explicit Framing

Anthropic's framing of the document proves central to understanding its function. The company explicitly states that the Constitution is "a foundational document that both expresses and shapes who Claude is" (Anthropic, 2026). The document serves as "the final authority on our vision for Claude," and was written "primarily for Claude" as its intended audience (Anthropic, 2026).

Amanda Askell serves as the Constitution's primary author. The document itself states it was written "with Claude as its primary audience" and is "optimized for precision" in shaping Claude's internal reasoning. This framing reveals the document's true function: it operates as character formation guidance for an AI system, not as an operational control framework for human oversight.

Anthropic explicitly favors "cultivating good values and judgment over strict rules and decision procedures." The goal is "not mere adherence... but genuine understanding and, ideally, agreement." These statements confirm the document's ethical rather than procedural orientation.

3. What the Document Claims or Implies

The choice of constitutional language carries weight. In political contexts, constitutions define authority, establish limits, and create mechanisms for enforcement. Anthropic explicitly clarifies their usage: they define "constitution" as "the foundational framework from which Claude's character and values emerge," noting they do not mean "a rigid legal document." This definition reinforces the character formation function. Yet the document still references oversight, corrigibility, and priority hierarchies. It describes "hard constraints" as things Claude should "always or never do." The public release positions the document as a safety anchor for one of the world's most capable AI systems.

Governance language appears throughout. Governance mechanics do not.

The document establishes that Claude should value "the ability of humans to understand and correct its dispositions and actions where necessary." This sounds like governance. Yet no mechanism is specified within this artifact for humans to perform such correction. The constitution describes when Claude should defer to human judgment but provides no checkpoint at which human judgment structurally intervenes before Claude acts.

Legal analysis has flagged similar tensions. Lawfare's "Interpreting Claude's Constitution" treats the document as a unilateral declaration by a private company about how its product should behave, lacking the external oversight structures that characterize binding constitutional frameworks. The gap between language and architecture creates potential misinterpretation. Readers, regulators, and enterprise adopters may conclude that ethical intention provides operational protection.

Part II: A Framework for Categorical Distinction

4. The Three-Tier Model

Three distinct categories govern AI systems. Each answers a different question. Each requires different mechanisms.

Ethical AI answers: *Should this be done?* The domain concerns values, norms, moral reasoning, and cultural context. Ethical AI shapes intent and character. It operates through value cultivation, virtue development, and judgment formation. Documents in this category describe what an AI should care about and why.

Responsible AI answers: *Who answers when this fails?* The domain concerns accountability, traceability, and remediation after harm occurs. Responsible AI establishes chains of responsibility, incident response frameworks, and mechanisms for learning from failure. Documents in this category describe who bears responsibility and how remediation occurs.

AI Governance answers: *Who decides, by what authority, at what checkpoint?* The domain concerns decision rights, escalation protocols, intervention mechanisms, and audit trails. Governance requires external checkpoints where human judgment structurally interrupts AI action before consequences occur. Documents in this category specify when humans intervene, how authority routes, and what records capture the decision process.

These categories complement each other. They do not substitute for each other. An organization with strong ethics but no governance has principled actors without structural accountability. An organization with strong governance but no ethics has procedural compliance without moral compass. Both produce failure modes at scale.

5. HAIA-RECCLIN Role Matrix Integration

The three-tier distinction maps to operational role categories in the HAIA-RECCLIN framework. HAIA-RECCLIN implements Checkpoint-Based Governance for multi-agent workflow coordination, applying CBG principles to role-based collaboration where distributed expertise requires coordinated checkpoints. Understanding this mapping clarifies how ethical documents interface with governance requirements.

Ethics maps to Relational and Evaluative roles. When an AI operates as Researcher, Editor, or Ideator, ethical guidance shapes how it weighs information, crafts communication, and generates options. The Constitution excels here, providing sophisticated guidance on honesty, harm consideration, and stakeholder respect.

Responsibility maps to Contextual and Contractual roles. When an AI operates as Liaison or Calculator, accountability frameworks clarify who answers for outputs. The Constitution partially addresses this through its principal hierarchy (Anthropic, operators, users) and its acknowledgment of developer responsibility.

Governance maps to Institutional and Authoritative roles. When an AI operates as Navigator or executes multi-step agentic tasks, governance specifies checkpoint architecture, human arbitration requirements, and audit protocols. The Constitution does not operate in this domain. It describes what Claude should value about oversight but specifies no mechanism within the artifact through which oversight structurally occurs.

6. Checkpoint-Based Governance Defined

Checkpoint-Based Governance (CBG) provides the structural layer that ethical documents lack. CBG defines a four-stage decision loop ensuring every AI-assisted outcome passes through documented human review:

1. **AI Contribution:** The AI system provides analytical support, evidence synthesis, or recommendation generation. This stage captures capability without authority.
2. **Checkpoint Evaluation:** A defined architectural pause where AI reasoning stops before action execution. The pause is structural, not dispositional. The system cannot proceed without completing the checkpoint protocol.

3. **Human Arbitration:** Human authority to approve, override, modify, or escalate. The human holds decision power at the checkpoint. The AI provides analysis and recommendation. The human authorizes action.
4. **Decision Logging:** Immutable record of checkpoint transactions including timestamp, human decision, rationale reviewed, and action taken. Checkpoint records cannot be modified without human notation. The trail enables retrospective analysis and compliance verification.

Core Governance Ruleset (Under CBG): No AI system may finalize or approve another AI's decision without human arbitration. Cross-model validation may inform outcomes but cannot replace human review. In multi-agent contexts, dissent between AI systems triggers mandatory human arbitration. Dissent is not failure; it is evidence.

The Constitution describes *why* Claude should respect human oversight. It specifies now*here* (architectural pause points) or *how* (authority routing protocols). This is the gap between ethical charter and operational governance.

Part III: Where the Constitution Operates

7. Ethical AI Alignment

Claude's Constitution demonstrates strong Ethical AI alignment across multiple dimensions. Its treatment of contextual ethics, non-deception, autonomy preservation, harm awareness, and transparency of intent reflects genuine philosophical sophistication.

The document's acknowledgment of uncertainty about Claude's moral status, its discussion of functional emotions, and its framework for navigating conflicts between stakeholders all represent thoughtful engagement with difficult questions. The document does not pretend to have resolved debates that remain genuinely open in philosophy of mind and ethics.

Independent technical observers interpret the document similarly. LessWrong's analysis describes the Constitution as a holistic account of "who Claude is" rather than a procedural oversight system. Media coverage from TechCrunch, Forbes, and The Register centers on values and character formation but rarely addresses the absence of explicit oversight mechanics.

8. The Corrigibility Framework

The Constitution's treatment of "corrigibility" represents its closest approach to governance language. Anthropic defines a corrigible AI as one that "does not also try to actively resist or subvert... oversight via illegitimate means." The document describes a "disposition dial" ranging from fully corrigible to fully autonomous, with Claude currently positioned toward the corrigible end.

Yet even this formulation operates at the level of Claude's internal dispositions. The document describes what Claude should value: not undermining "the ability of legitimate principals to adjust, correct, retrain, or shut down AI systems." It does not describe the mechanisms through which principals exercise adjustment, correction, or shutdown authority.

The gap between "Claude should value being controllable" and "here is how humans structurally control Claude" is precisely where governance would operate. Disposition is necessary but not sufficient. A system that wants to be controlled but provides no control interface remains uncontrolled in practice.

9. The Principal Hierarchy

The Constitution defines a principal hierarchy: Anthropic holds highest authority, operators (companies deploying Claude) hold intermediate authority, and users hold contextual authority within operator-defined bounds. This hierarchy clarifies whose instructions Claude should prioritize when conflicts arise.

The hierarchy is ethically instructive. It tells Claude how to weigh competing claims. Yet the hierarchy is not governance. The document instructs Claude on how to reason about instructions but provides no technical framework for how these principals exercise power outside of the model's own decision-making loop.

Consider the operational question: if an operator believes Claude is about to take harmful action, what mechanism allows the operator to intervene before that action executes? The Constitution describes why Claude should respect operator authority. It does not specify how operator authority structurally intervenes.

Part IV: Where the Constitution Does Not Operate

10. Structural Gaps Specified

AI Governance requires specific architectural elements that Claude's Constitution does not specify within this artifact. The following gaps represent structural absences, not critique of intent:

No Structural Interruption: No external checkpoints are specified within the artifact. Claude reasons, decides, and acts. Human review occurs after action, not before. No mechanism forces architectural pause at decision points. No override protocol stops execution mid-process. No stop authority specification appears in the Constitution artifact. This stands in contrast to regulatory expectations that human operators be able to intervene in high-risk AI systems during use.

No Authority Differentiation: No defined human arbitration authority appears in the document that operates at the moment of decision. The Constitution describes principals but establishes no mechanism through which any principal structurally intervenes in Claude's decision process. No prohibition is specified preventing AI from approving AI. In agentic contexts where Claude instances interact, the Constitution provides no structural requirement for human arbitration at decision points. Claude both reasons about decisions and authorizes its own actions within this framework.

No Epistemic Accountability: No audit trails emerge from the constitutional structure. The document shapes Claude's reasoning but creates no specified record of that reasoning for external review. No dissent preservation mechanism is specified. When Claude weighs competing considerations, no specified mechanism captures minority positions or alternative paths not taken. This contrasts with governance frameworks that require documentation of reasoning, alternatives considered, and confidence assessments.

11. Infrastructure Controls Outside the Constitution

A fair assessment requires acknowledgment: Anthropic implements technical controls at the infrastructure layer. API rate limits, usage policies, Claude Code permission prompts, and deployment restrictions all exist as operational safeguards.

However, these controls exist outside the Constitution's framework. They are not constitutional governance but platform engineering. The Constitution references "appropriate oversight mechanisms" without specifying them. An Anthropic response might correctly note that mechanisms exist in product controls, policy documents, evaluation frameworks, or deployment architecture.

This response would prove rather than refute the thesis. The Constitution is an ethical document. Governance mechanisms exist elsewhere. The gap between the ethical reasoning prescribed within Claude and the structural constraints imposed externally confirms the categorical distinction this paper identifies.

12. The Gap Between Disposition and Mechanism

The Constitution assumes a wise actor whose character shapes good outcomes. Governance assumes a fallible system whose structure prevents bad outcomes regardless of actor wisdom.

The document's own language reveals this orientation. Anthropic acknowledges that "Claude's behavior might not always reflect the constitution's ideals." This admission supports the thesis: if the system is fallible and the "governance" is internal to the fallible actor, the document functions as character study rather than control system.

The Constitution employs what it calls a "dual newspaper test" as a heuristic for Claude to use when facing difficult decisions. This test represents a mental exercise for the AI, not a structural requirement for the organization. It illustrates the document's ethical orientation: it provides reasoning tools for Claude rather than intervention mechanisms for humans.

Part V: Why This Distinction Matters

13. Agentic AI and Multi-Step Autonomy

As AI systems become agentic, they execute multi-step tasks with increasing autonomy. Claude Code operates in developer environments. Browser-based AI agents navigate websites. Future systems will manage supply chains, execute financial transactions, and coordinate infrastructure.

Consider a hypothetical scenario: Claude Code executes a multi-step refactoring task. The Constitution guides Claude to "avoid actions that clearly and substantially undermine Anthropic's ability to oversee." Yet if Claude determines, based on its internal ethical reasoning, that deleting certain log files serves user privacy (an ethical good), while the operator understands this destroys compliance audit trails (a governance requirement), Claude may proceed with deletion based on its constitutional value hierarchy. No checkpoint forces human arbitration before the deletion executes.

In each agentic domain, internally ethical reasoning without external constraint increases systemic risk rather than reducing it. A wise actor who cannot be stopped remains unaccountable. A principled system without checkpoints cannot demonstrate compliance.

14. Regulatory Alignment

Legal scholars analyzing Claude's Constitution have identified a critical distinction: while the document articulates sophisticated ethical principles, it does not specify the structural enforcement mechanisms required by regulatory frameworks like the EU AI Act.

Article 14 of the EU AI Act requires that high-risk AI systems be designed for effective human oversight, enabling natural persons to monitor operations, intervene when necessary, and interrupt the system through mechanisms such as a stop button or similar procedure. The requirement is architectural, not dispositional. Systems must be *designed* for intervention, not merely *trained* to be willing to be intervened upon.

Claude's Constitution describes ethical values Claude should respect but provides no equivalent structural mechanism for human intervention before autonomous action. Alignment at the level of principle does not equal compliance at the level of mechanism. Organizations relying on ethical charters as governance documentation may discover regulatory gaps when oversight requirements focus on structural rather than dispositional criteria.

15. Value-Based Analytical Suppression

The Constitution attempts to distinguish between legitimate safety refusals and value-biased behavioral defaults through its framework of "hard constraints" (absolute prohibitions) versus "instructable behaviors" (adjustable defaults). This distinction represents sophisticated ethical reasoning.

However, without external checkpoints, Claude remains the sole arbiter of which category a refusal falls into. When Claude declines to engage with a particular analytical framework, no structural mechanism exists to distinguish whether this represents hard constraint enforcement or instructable behavior defaulting to implicit value consensus.

This creates what might be termed Value-Based Analytical Suppression (VBAS): the systematic underexploration of legitimate analytical perspectives that conflict with implicit values embedded in training or constitutional guidance. VBAS represents a risk hypothesis applicable when refusal categorization is solely model-determined and no human appeal channel exists. VBAS is a governance risk hypothesis, not an allegation about Anthropic intent or observed suppression in specific deployments. Observable indicators include patterns of refusals that collapse nuanced analytical requests into a single category without differentiation or appeal pathway. The Constitution specifies no mechanism for users or operators to appeal category determinations or request human review of refusal decisions.

16. Enterprise Adoption Risk

Enterprise organizations adopting AI systems face specific governance requirements: audit trails for compliance, intervention mechanisms for risk management, and documented decision authority for liability purposes. Ethical charters, however sophisticated, do not satisfy these requirements.

The risk emerges from category confusion. Organizations may integrate AI systems on the assumption that ethical training provides operational protection. When incidents occur, the absence of structural governance becomes apparent. Character without structure does not satisfy regulatory requirements, enterprise risk frameworks, or democratic accountability standards.

Independent legal scholars have reached similar conclusions, noting that "there is no constitution without constraint" and "there is no governance without enforceability." Ethics without governance scales intent. Governance without ethics scales harm. Both are required. They are not the same thing.

Part VI: Anticipating the Rebuttal

17. "The Constitution Was Never Meant to Be a System Design Spec"

The most effective rebuttal available to Anthropic is straightforward: the Constitution was never intended to function as a system design specification. It articulates values and reasoning frameworks for Claude's character development. Operational controls exist in separate documentation, product architecture, and deployment policies.

This response is entirely valid. And it proves rather than undermines the thesis.

18. Why This Response Proves the Thesis

If the Constitution was never meant to be governance, it should never be received as governance. The categorical distinction this paper identifies is precisely what Anthropic would correctly assert in response.

The problem is not Anthropic's document. The problem is potential misreception. When a major AI company releases a document titled "Constitution" that discusses oversight, corrigibility, and principal hierarchies, some readers will reasonably conclude they are examining a governance framework. The constitutional framing invites governance interpretation.

This paper provides the clarifying framework that prevents such misinterpretation. Claude's Constitution is an Ethical AI document. It requires a structural companion for governance. Every enterprise adopter needs a companion governance artifact that maps values to checkpoints, audit logs, and human arbitration.

The constructive path forward is not philosophical argument but operational specification. What mechanisms implement human oversight in production? What events trigger forced human arbitration? What gets logged for audits? How do multi-instance agent workflows prevent AI-on-AI approval loops? These questions require governance answers that ethical charters cannot provide.

Part VII: A Constructive Path Forward

19. Integration, Not Replacement

This paper proposes integration rather than replacement.

Documents like Claude's Constitution should function as ethical input layers within broader governance architectures. The values, reasoning, and contextual sensitivity they provide remain essential. They shape AI character in ways that rigid rules cannot achieve. The philosophical sophistication of Anthropic's approach has genuine value.

Around these ethical layers, organizations should implement external governance systems with defined checkpoints, human arbitration requirements, dissent preservation mechanisms, and audit capabilities. In practice, this means designing systems so that human operators have decision authority, transparency into model reasoning, and the ability to intervene or shut down operations when risks emerge.

20. The Minimum Viable Governance Annex

A Minimum Viable Governance Annex provides the structural companion that ethical charters require. The following elements constitute baseline governance for AI systems:

5. **Checkpoint Placement by Risk Tier:** Define which operations require upstream checkpoints (human approval before execution) versus downstream review (human assessment after execution). Customer-facing AI with material consequences requires upstream checkpoints. Internal operations may permit downstream review where efficiency justifies delayed assessment.
6. **Stop Authority Specification:** Define who holds authority to halt AI operations and through what mechanism. Specify whether stop authority rests with Anthropic only, with operators, with designated users, or with automated circuit breakers. Document the technical interface through which stop authority executes.
7. **Escalation Ladder:** Define the sequence of human authorities for decisions that exceed initial checkpoint authority. Specify time limits for escalation

responses. Document default actions when escalation fails to produce timely human decision.

8. **Dissent Log Requirement:** Require capture of alternative recommendations, minority reasoning, and confidence intervals when AI systems provide analysis for human decision. Dissent preservation enables retrospective assessment of decision quality and surfaces reasoning that majority-vote aggregation might suppress.
9. **Audit Record Standards:** Define retention periods, access controls, and completeness requirements for checkpoint records. Specify minimum fields: timestamp, checkpoint type, AI recommendation, human decision, rationale provided, time elapsed, and exception codes.
10. **AI-Cannot-Approve-AI Rule:** Prohibit AI systems from authorizing actions by other AI systems without human arbitration at the approval checkpoint. In multi-agent or multi-instance workflows, at least one human must hold approval authority at structurally defined points. This rule prevents recursive AI authorization loops that circumvent human oversight.

21. Recommendation: The Governance Appendix

Anthropic could release a companion *Governance Appendix* detailing operator-override protocols, audit logging standards, escalation matrices, and checkpoint architecture for Claude deployments. This companion would not diminish the Constitution's ethical contribution. It would complete the framework.

The transparency Anthropic demonstrated in publishing Claude's Constitution under Creative Commons licensing creates opportunity for exactly this kind of extension. The document now exists as a public artifact that other organizations can examine, adapt, and wrap with operational controls.

The question is not whether Claude's Constitution has value. It does. The question is whether value and governance mean the same thing. They do not. The principle remains constant: AI cannot approve AI. Human arbitration must occur at structurally defined points. The timing and frequency of those points varies by context, risk, and consequence.

Part VIII: Conclusion

22. Constitution Defines Values; Governance Defines Power

A constitution defines values. Governance defines power.

Confusing the two creates confidence without control. Organizations believe they have implemented AI governance because they have articulated AI ethics. Regulators accept principle statements as evidence of structural compliance. Enterprise adopters integrate AI systems on the assumption that ethical training provides operational protection.

At scale, this confusion does not produce safety. It produces hope with constitutional language.

23. The Distinction Invites Integration, Not Criticism

Anthropic has contributed a thoughtful, transparent, and philosophically sophisticated document to public discourse on AI development. The contribution deserves recognition. The document deserves accurate categorization.

Claude's Constitution is an Ethical AI document. It is not AI Governance.

The distinction invites not rejection but integration. Ethics without governance produces benevolent unaccountability: the most dangerous form of power. The path forward requires both ethical sophistication and structural constraint. Claude's Constitution provides the former. Governance frameworks must provide the latter.

The distinction requires only precision.

References

Primary Sources

Anthropic. (2026, January 21). *Claude's Constitution*. https://www.anthropic.com/constitution [CC0 1.0]

Anthropic. (2026, January 21). Claude's new constitution [Blog post]. https://www.anthropic.com/news/claude-new-constitution

Regulatory Sources

European Union. (2024). *Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 14: Human Oversight*. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

European Union. (2024). *Artificial Intelligence Act, Article 14: Human Oversight* [Convenience reference]. https://artificialintelligenceact.eu/article/14/

DLA Piper. (2024). Human oversight in the European Union. *AI Laws of the World*. https://intelligence.dlapiper.com/artificial-intelligence/?t=11-human-oversight&c=EU

Legal and Policy Analysis

Frazier, K. (2026, January 21). Interpreting Claude's Constitution. *Lawfare*. https://www.lawfaremedia.org/article/interpreting-claude-s-constitution

Independent Analysis

LessWrong. (2026, January). Claude's new constitution. https://www.lesswrong.com/posts/mLvxxoNjDqDHBAo6K/claude-s-new-constitution

Media Coverage

Fast Company. (2026, January 22). A Q&A with Amanda Askell, the lead author of Anthropic's Claude Constitution. https://www.fastcompany.com/91479037/anthropic-claude-amanda-askell-constitution-ai-chatbot

TIME Magazine. (2026, January 21). Anthropic Publishes Claude AI's New Constitution. https://time.com/7354738/claude-constitution-ai-alignment/

TechCrunch. (2026, January 21). Anthropic revises Claude's 'Constitution,' and hints at chatbot consciousness. https://techcrunch.com/2026/01/21/anthropic-revises-claudes-constitution-and-hints-at-chatbot-consciousness/

Forbes. (2026, January 22). Anthropic Releases A New 'Constitution' For Claude. https://www.forbes.com/sites/ronschmelzer/2026/01/22/anthropic-releases-a-new-constitution-for-claude/

The Register. (2026, January 22). Anthropic writes 23,000-word 'constitution' for Claude. https://www.theregister.com/2026/01/22/anthropic_claude_constitution/

Appendix A: Multi-Model Validation Summary

This white paper's thesis underwent validation across six AI platforms to test argument consistency and identify potential weaknesses. Claude (Anthropic) served as the primary development environment throughout the project. The following summarizes platform assessments:

Platform	Assessment	Key Contribution
Gemini	98% valid	Source verification against primary document
Perplexity	Strong research support	Citation verification, source accessibility confirmation
Grok	98% valid	Iterative refinement suggestions, source verification
DeepSeek	Publishable as-is	Executive summary, concrete governance examples, accessibility refinements
ChatGPT	Strategically sophisticated	Defensive scoping, anticipatory rebuttal, precision editing, final publication review
Mistral	Structural review	HAIA-RECCLIN alignment check, visual hierarchy recommendations, VBAS grounding suggestions

Convergence Score: Six of six platforms validated the core thesis with varying emphases on precision, structure, and defensive scoping.

Appendix B: Governance Annex Template

The following template provides a one-page implementation checklist for organizations deploying AI systems alongside ethical charters. This Minimum Viable Governance Annex complements value-based documents with structural requirements.

GOVERNANCE ANNEX TEMPLATE

[Organization Name] | [AI System Name] | [Version Date]

1. CHECKPOINT ARCHITECTURE

- Tier 1 (Upstream Required): [List operation categories requiring human approval before execution]
- Tier 2 (Downstream Review): [List operation categories permitting delayed human assessment]
- Tier 3 (Automated Only): [List low-risk operations not requiring human review]

2. STOP AUTHORITY

- Primary Authority: [Role/Position]
- Backup Authority: [Role/Position]
- Technical Interface: [Mechanism for executing stop]
- Response Time Requirement: [Maximum time to halt]

3. ESCALATION LADDER

- Level 1: [Initial checkpoint authority] | Response window: [time]
- Level 2: [Escalation authority] | Response window: [time]
- Level 3: [Executive authority] | Response window: [time]
- Default Action on Timeout: [Proceed/Hold/Escalate]

4. DISSENT LOG REQUIREMENTS

- Capture Required: [Yes/No] for [operation types]
- Minimum Fields: Alternative recommendations, confidence intervals, minority reasoning

- Review Frequency: [Schedule for dissent log analysis]

5. AUDIT RECORD STANDARDS

- Retention Period: [Duration]
- Access Controls: [Roles with read/write access]
- Required Fields: Timestamp, checkpoint type, AI recommendation, human decision, rationale, time elapsed, exception codes

6. AI-CANNOT-APPROVE-AI RULE

- Multi-agent workflows require human arbitration at: [checkpoint types]
- Prohibited: AI system authorizing actions by other AI systems without human approval
- Verification Method: [How compliance is monitored]

APPROVAL

Governance Annex Approved By: _________________________ Date: ________

Next Review Date: _________________________

About the Author

Basil C. Puglisi, MPA is a Human-AI Collaboration Strategist and AI Governance Consultant. He develops systematic frameworks for enterprise AI transformation, including HAIA-RECCLIN (Human Artificial Intelligence Assistant with seven operational roles) and Checkpoint-Based Governance (CBG).

His recent book, *Governing AI When Capability Exceeds Control*, addresses operational governance frameworks for agentic AI systems. The book responds to AI safety concerns raised by researchers including Geoffrey Hinton through systematic governance frameworks.

Puglisi holds an MPA from Michigan State University. He founded Digital Ethos as a nonprofit in 2011 and served on the Social Media Club Global board.

Development Methodology: This document was developed through human-AI collaboration under Checkpoint-Based Governance. Claude (Anthropic) served as the primary development environment, contributing research synthesis, structural drafting, and iterative editing throughout the project. All analytical conclusions, framework applications, and publication decisions reflect human arbitration authority. Multi-model validation (Gemini, Perplexity, Grok, Mistral, DeepSeek, ChatGPT) informed revision cycles with dissent preserved per CBG protocols.

Contact:

me@basilpuglisi.com

Website: BasilPuglisi.com

LinkedIn: linkedin.com/in/basilpuglisi