POSITION PAPER

The Missing Governor: Anthropic's Constitution and Essay Acknowledge What They Cannot Provide

A Structural Response to Claude's Constitution & "The Adolescence of Technology" Essay

Basil C. Puglisi, MPA

Human-AI Collaboration Strategist

BasilPuglisi.com

January 2026

Version 2.3 | January 2026 | HAIA-RECCLIN Validated (Claude, ChatGPT, Perplexity, Grok, Mistral, Deepseek, Kimi & Gemini) | Checkpoint Based Governance by Basil C. Puglisi

Executive Summary

On January 21, 2026, Anthropic published Claude's Constitution, an 80-page document articulating values, character formation, and behavioral guidelines for its AI system. Six days later, on January 27, 2026, CEO Dario Amodei released "The Adolescence of Technology," a 20,000-word essay examining AI risk and calling for societal response. The timing was not coincidental. Together, these publications represent a coordinated statement on AI development philosophy from the company holding 32% of enterprise LLM market share.

This paper examines what the combined publications address, what they acknowledge but leave architecturally undefined, and proposes a framework for filling the remaining gaps.

The analysis reveals that Anthropic's combined position comprehensively addresses two critical layers: Ethical AI (character formation, values, honesty properties) and Responsible AI (training methodology, internal safeguards, corrigibility architecture). Both documents acknowledge the need for external governance. Amodei explicitly calls for legislation. The Constitution references "legitimate external factors like government regulation." However, neither document provides architecture for how such external governance would operate.

This paper establishes a foundational principle: AI Governance requires a Human Governor, and no machine completes governance regardless of sophistication. This distinction is definitional rather than preferential. The Human Governor stands accountable through moral, employment, civil, and criminal channels, and this accountability creates the incentive to do better, to be ethical, to be thorough. That incentive structure is built into the threat facing humans but absent from machines. AI Governance is categorically higher than Responsible AI because someone can be held to account.

The paper proposes Checkpoint-Based Governance (CBG) as framework for the missing governance layer. CBG does not replace what Anthropic has built. It proposes an external complement that addresses the governance gap both documents acknowledge.

This paper's distinctive contribution is the epistemic coverage argument: no individual, regardless of credentials, can inhabit experiential positions they have not occupied. Constitutional committees with diverse experiential knowledge are the structural solution to value formation authority. Amanda Askell, the philosopher responsible for Claude's constitution, herself acknowledges this limitation, stating in her January 2026 Vox interview: "I'm thinking about this a lot. And I want to massively expand the ability that

we have to get input." This paper proposes architecture that addresses the limitation Anthropic's own constitutional author recognizes.

1. The Human Governor Principle

AI Governance requires a Human Governor. No AI system, however sophisticated, completes governance independently. This is the foundational principle from which all else follows.

The machine has no incentive built into threat. You cannot threaten a machine with moral judgment. You cannot threaten a machine with termination. You cannot sue a machine. You cannot imprison a machine. The machine processes inputs and generates outputs with complete indifference to consequence.

The human faces consequence, and that consequence creates the incentive to do better, to be careful, to be ethical, to be thorough. The entire structure surrounding human action builds this incentive in.

1.1 The Accountability Architecture

Moral Accountability as Incentive. The Human Governor knows their judgment will be evaluated by peers, profession, and community. Reputation is at stake, professional standing is at stake, and the Governor who cuts corners, who rubber stamps, who fails to engage seriously knows that their standing depends on doing the work.

Employment Accountability as Incentive. The Human Governor knows their job depends on quality of judgment. Poor performance leads to remediation, reassignment, termination. The Governor who wants to remain employed has incentive to govern well.

Civil Accountability as Incentive. The Human Governor knows that negligent judgment may result in lawsuit, personal liability, and financial consequence that affects their life, their family, their future.

Criminal Accountability as Incentive. The Human Governor knows that gross negligence or recklessness may result in prosecution, with liberty itself at stake. Even Governors tempted to cut corners recognize the ultimate backstop: if you are reckless enough, society will take your freedom.

This incentive structure does not exist for machines. Responsible AI operates without any system component that experiences threat and modifies behavior to avoid it. Quality systems, monitoring, constitutional training, safeguards: all of these are engineering choices made by humans, not incentive responses by the system itself. The machine does not try harder because it fears nothing.

1.2 Why AI Governance Is Categorically Higher

AI Governance is categorically higher than Responsible AI, and the reason is not that humans are more accurate but that accountability creates incentive.

The need to do better, to be ethical, to be thorough, to care about outcomes is built into the threat structure facing humans. The execution is imperfect because humans still fail, still cut corners, still act negligently, but the incentive exists and the structure pushes toward better behavior because worse behavior carries consequence.

This is not denigrating Responsible AI but recognizing what accountability provides. Responsible AI without human checkpoint operates on engineering quality alone, while AI Governance with human checkpoint operates on engineering quality plus human incentive to get it right because getting it wrong carries personal consequence.

The EU AI Act Article 14 validates this categorical distinction through binding legal force. Article 14 mandates that high-risk AI systems be designed so they "can be effectively overseen by natural persons during the period in which the AI system is in use." The regulation specifies four oversight models: Human-in-Command with ultimate veto authority, Human-in-the-Loop with direct operational involvement, Human-on-the-Loop with supervisory oversight, and Human-over-the-Loop with strategic governance. Enforcement beginning August 2026 carries penalties reaching EUR 35 million or 7% of global revenue. This regulatory recognition confirms that governance without accountable humans is categorically insufficient.

1.3 Scope Clarification

The Human Governor Thesis defines AI Governance (Handmade Quality). Organizations operating in Responsible AI mode (Factory Quality) accept that AI validates AI. The thesis applies when stakes require governance; it does not mandate governance for all contexts. Choosing Responsible AI over AI Governance is a legitimate organizational decision based on risk tolerance and consequence severity.

2. The Three-Layer Framework

Three distinct concepts often collapse into a single term, creating confusion that serves no stakeholder. These layers separate cleanly when articulated with precision.

2.1 Ethical AI: The Values Line

Ethical AI establishes foundation through value formation, character architecture, and constitutional grounding. All AI begins here at the values line: what should the system embody? How should it reason about moral questions? What character traits should training cultivate?

This layer concerns the system's fundamental orientation toward beneficial behavior, independent of external enforcement. Anthropic's Constitution addresses this layer comprehensively, articulating seven honesty properties: truthful, calibrated, transparent, forthright, non-deceptive, non-manipulative, and autonomy-preserving.

The Constitution emphasizes "cultivating good values and judgment over strict rules and decision procedures." This represents a sophisticated approach to character formation that recognizes rigid rules cannot anticipate every situation. Amanda Askell explicitly describes her approach as virtue ethics in her January 2026 Vox interview. This maps precisely to the Ethical AI category: character formation rather than rule-based systems.

2.2 Responsible AI: The Accountability Line

Responsible AI evolves from Ethical AI through natural progression as character formation leads to technical implementation, values become safeguards, and constitutional principles become monitoring systems.

This is the accountability line: disclosure, sourcing discipline, editor and publisher expectations, and who answers when harm shows up. How do developers translate ethical aspirations into technical reality? What training approaches, safety mechanisms, and operational constraints ensure systems behave as intended?

Responsible AI has a ceiling, and that ceiling is the absence of individual human oversight. Even sophisticated Responsible AI with random spot-checking remains machine checking machine where algorithms verify algorithms and systems audit systems. Humans review aggregates, trends, and exceptions, but no human stands accountable for individual outputs.

This ceiling defines the category. Responsible AI can become extraordinarily capable. It can implement constitutional values with precision. It can monitor itself with rigor. What it cannot do is place a human in the decision chain for individual outputs.

2.3 AI Governance: The Authority Line

AI Governance requires transformation of the fundamental relationship between human judgment and individual outputs, with a Human Governor present for outputs that matter. Each governed output receives individual human attention before consequence.

This is the authority line: the checkpoints that decide what gets used, labeled, rejected, and what stays human only. Who holds binding power to approve, modify, or halt AI deployment? What mechanisms ensure those authorities can exercise their power effectively? What happens when the entity building the system disagrees with external judgment?

The pathway from Responsible AI to AI Governance is open. Adding genuine human oversight on individual outputs transforms the category. But adding checkpoints to factory operations does not make them handmade. Factory with quality control remains factory with quality control. The quality distinction is irreducible.

2.4 The Grammar Distinction

Notice the grammar: Ethical AI, Responsible AI, AI Governance. In the first two, AI sits as the noun, and ethics or responsibility modifies the machine. In governance, the structure reverses. AI modifies governance, and the human system holds the final position. This reflects where authority lands.

The sequence is temporal, not just conceptual, because ethics precedes design, Responsible AI operates during design and deployment, and governance operates after outputs. Each layer has a different locus of control.

2.5 The Terminological Boundary

Responsible AI must never be called AI Governance without direct human oversight.

This is not semantic preference but definitional integrity because governance implies authority, authority implies accountability, and accountability requires an accountable party. Machine checking machine produces no accountable party. When outputs cause harm, who answers? The algorithm? The training data? The constitutional principles? None of these can be held accountable in ways that matter.

The term "AI Governance" is reserved for systems where a human stands in the decision chain, exercising judgment, bearing responsibility, subject to consequence.

3. The Quality Distinction

Adding checkpoints to factory operations does not make them handmade. Factory with quality control remains factory with quality control. The quality distinction is irreducible.

3.1 Factory Production (Responsible AI)

Mass production at scale with quality control through sampling and automated inspection produces output that is efficient, affordable, and widely available. Most outputs are fine, and the consumer knows this was factory process where errors are possible, not likely, but possible.

Responsible AI with Checkpoints adds human review at defined intervals or for flagged cases. Still fundamentally factory with quality control stations. The checkpoint improves reliability. It does not transform each output into handmade work. This is enhanced Responsible AI, not AI Governance.

Some AI will remain at Responsible AI permanently because of scale. This is not failure. This is appropriate placement based on the nature of the application. Consumer chatbots processing millions of interactions, recommendation engines serving billions of suggestions, content moderation at platform scale: these operate appropriately under Responsible AI with honest disclosure.

3.2 Handmade Production (AI Governance)

Human checkpoints throughout the process mean someone inspected this specific item before it reached you, not through sampling or statistical quality control but through individual attention. Errors become extremely unlikely because humans were present at points where those errors would be caught.

The test is not whether humans are involved. The test is whether human judgment is present for the outputs that matter.

3.3 The Authorship Test

If you cannot defend an AI output line by line to a skeptical expert, you have not authored it but merely forwarded it. Authorship requires understanding, not just approval. The Human Governor reviewing military targeting decisions must be able to explain and defend those decisions. The author publishing AI-assisted writing must be able to explain and defend that writing. The standard scales from life-and-death governance to craft and provenance in creative work. The checkpoint holds because the human behind it can account for what passed through.

3.4 The Moral Foundation

One innocent death is too many, and this is not rhetorical but the principle that determines when Responsible AI, however sophisticated, is categorically insufficient.

99.9% accuracy sounds impressive. In a system processing one million decisions, 99.9% accuracy produces one thousand failures. If those decisions concern life and death, 99.9% accuracy produces one thousand deaths. No statistical success rate makes those deaths acceptable.

But we are nowhere near 99.9% accuracy now. Current AI systems hallucinate, fabricate sources, generate confident errors, and fail in ways their operators do not anticipate. Deploying such systems without qualified human oversight is not innovation but unprofessional conduct, and in domains with meaningful stakes, it is negligence.

The 99.9% discussion is academic distraction because current AI reliability mandates governance now. Failure to establish AI Governance over AI use is currently unprofessional at best, negligent at worst. There must be humans who are qualified to conduct AI governance and who can be held accountable, and the question is not when governance becomes necessary because the necessity already exists.

Medical device regulation demonstrates governance necessity when stakes involve life. FDA Premarket Approval for Class III devices requires human checkpoint authority at four stages: administrative review for completeness, in-depth scientific and regulatory assessment, advisory panel review by independent committee, and final FDA decision. No manufacturer self-certification suffices. External binding authority reviews every high-risk device before deployment. The checkpoint architecture scales to consequence: Class I devices receive light oversight through general controls, Class II devices require 510(k) clearance demonstrating substantial equivalence, Class III devices affecting life receive governance regardless of manufacturer preference. AI governance follows established regulatory precedent, not untested philosophy.

4. What Anthropic's Documents Address

4.1 Ethical AI: Comprehensive Coverage

The Constitution thoroughly addresses character formation. It articulates that Claude should have "good values," be "honest," and maintain "broadly ethical" behavior. The document lists seven honesty properties: truthful, calibrated, transparent, forthright, non-deceptive, non-manipulative, and autonomy-preserving.

The document establishes a clear priority hierarchy: (1) being safe and supporting human oversight, (2) behaving ethically, (3) following Anthropic's guidelines, and (4) being helpful. If Claude is conflicted, Anthropic wants the model to "generally prioritize these properties in the order in which they are listed."

The Constitution distinguishes between hardcoded behaviors (absolute prohibitions such as providing bioweapons assistance or generating child sexual abuse material) and softcoded defaults that operators and users can adjust within defined boundaries.

4.2 Responsible AI: Detailed Implementation

The Constitution details internal safeguards with remarkable specificity. The "principal hierarchy" establishes trust relationships: Anthropic holds highest authority, operators receive intermediate trust, users receive baseline trust. "Corrigibility" emerges as a key safety property, meaning Claude should remain controllable and correctable by appropriate authorities.

Amodei's essay discusses complementary technical mechanisms: mechanistic interpretability for understanding model internals, self-monitoring capabilities, and staged autonomy increases tied to demonstrated trustworthiness. He writes: "We believe that a feasible goal for 2026 is to train Claude in such a way that it almost never goes against the spirit of its constitution."

Together, these publications present a thorough Responsible AI framework.

4.3 The Governance Gap Acknowledged

Both documents acknowledge the need for external governance. Amodei explicitly calls for legislation:

> *"A credible risk of danger is enough for me and for Anthropic to pay quite significant costs to address it, but once we get into regulation, we are forcing a wide range of actors to bear economic costs."*

The Constitution references "legitimate external factors like government regulation" as constraints Claude should respect.

However, neither document provides architectural details for how such external governance would operate. The Constitution establishes internal principal hierarchy, which constitutes internal governance. Internal control is real, but it stays voluntary because the builder remains the final authority. External governance begins when someone outside the builder can stop the system and that stop survives commercial pressure. What remains architecturally undefined is external governance: binding oversight mechanisms, third-party veto power, multi-stakeholder decision authority, checkpoint architectures that survive the removal of voluntary cooperation.

This gap is not criticism of what Anthropic has done. It is recognition of what remains to be built. Amodei does not claim Anthropic has solved governance. He claims Anthropic has addressed internal mechanisms while explicitly calling for external mechanisms that others must build.

4.4 Addressing the Infrastructure Counterargument

Some argue that mechanisms exist outside the Constitution: API limits, permission prompts, usage policies, abuse detection systems, and ISO/IEC 42001 compliance. ISO 42001 is the international standard for AI management systems, providing structured methodology through Plan-Do-Check-Act cycles. It specifies "requirements for establishing, implementing, maintaining, and continually improving an AI management system" within an organization. These represent legitimate Responsible AI practice, but they do not constitute governance.

ISO 42001 is explicitly a management system standard focused on organizational internal processes. Management systems, however rigorous, remain self-governance. This is definitional category error when the question is external accountability. Governance requires authority independent of the entity being governed.

Infrastructure controls are engineering choices made by the same entity that built the system. They lack the essential properties of governance: external authority, binding enforcement, independent audit, and accountability to parties outside the commercial ecosystem.

The Constitution explicitly notes it does not apply to all contacts, such as some military contracts. This selective application demonstrates that constitutional principles are organizational policy, not governance architecture. Governance cannot be waived by the governed entity.

5. Why This Distinction Matters

5.1 Agentic AI and Multi-Step Autonomy

The stakes escalate with agentic AI. Claude Code, released in February 2025 and made generally available in May 2025, enables developers to delegate coding tasks directly from their terminal. The AI performs multi-step operations, accesses file systems, and executes code with increasing autonomy.

In multi-step agentic workflows, errors compound. A mistake at step three affects all subsequent steps. Constitutional training shapes character, but character alone cannot prevent compounding failures in autonomous operation. Governance requires intervention points where human judgment can intercept errors before they cascade. A checkpoint is a forced pause before an irreversible action, with a named human decision and an immutable record.

The Constitution addresses Claude's disposition toward helpful and safe behavior. It does not specify checkpoint architecture for agentic workflows. When Claude Code modifies production systems, who holds authority to halt the operation mid-stream? The Constitution shapes Claude's inclination to behave safely. External governance would define who can override Claude's execution and under what conditions.

5.2 Regulatory Alignment: EU AI Act Article 14

The EU AI Act Article 14 requires human oversight for high-risk AI systems. The Constitution's four-tier priority hierarchy aligns with EU AI Act requirements, as noted by the Bloomsbury Intelligence and Security Institute. This alignment positions Claude favorably for adoption by regulated industries.

But alignment with requirements differs from satisfaction of requirements. Human oversight under Article 14 means humans with authority to intervene, not AI systems trained to value human oversight. The Constitutional statement that Claude should support human oversight expresses disposition. The regulatory requirement demands mechanism.

Full EU AI Act enforcement begins August 2026, with penalties reaching EUR 35 million or 7% of global revenue. Organizations deploying Claude in high-risk contexts will need to demonstrate governance architecture, not merely constitutional values. The Constitution is a valuable starting point. It is not a compliance destination.

5.3 Value-Based Analytical Suppression (VBAS)

When AI systems are trained to embody certain values, those values shape what the system surfaces and what it suppresses. A system trained to avoid harmful content may suppress legitimate analysis of harmful phenomena. A system trained toward consensus may suppress legitimate minority positions.

This behavior is not malfunction but rather the logical consequence of value-based training applied to analytical tasks. The system behaves exactly as trained, and the suppression occurs precisely because the training succeeded.

Governance provides the counterweight. When humans hold decision authority at checkpoints, suppressed analysis can be surfaced through explicit request. Minority positions can be preserved. Uncomfortable findings can be examined. The human arbiter can override value-based filtering when the analytical task requires it.

Without governance architecture, value-based suppression operates without check. The Constitution acknowledges that Claude should preserve human autonomy and avoid manipulation. Governance would provide the structural mechanism ensuring these dispositions translate into actual practice when the system's trained values might otherwise filter relevant information.

5.4 Enterprise Adoption Risk

Enterprise adopters face a specific risk: confidence without control. The Constitution's sophistication inspires confidence. Its philosophical rigor suggests maturity. Its transparency demonstrates good faith.

Yet confidence is not control, and enterprise compliance requirements, particularly in healthcare, financial services, and government, demand documented human oversight. They demand audit trails showing who approved what and accountability chains traceable to individual humans. The audit asks for named roles with stop authority, the escalation window, the log schema, and the evidence that the stop works under incident pressure.

The Constitution shapes Claude's character. Enterprise governance requires checkpoint architecture independent of Claude's character. These are complementary needs. Addressing one does not address the other.

Fortune reports that Anthropic holds 32% of enterprise LLM market share by usage. Enterprise customers choosing Claude do so in part because of its safety reputation. That reputation creates obligation: enterprises trusting Claude's constitutional approach deserve clarity about where constitutional values end and where external governance must begin.

6. Constitutional Committee Design: The Epistemic Coverage Argument

This paper's distinctive contribution is the epistemic coverage argument: no individual, regardless of credentials, can inhabit experiential positions they have not occupied. Constitutional committees with diverse experiential knowledge are the structural solution.

6.1 The Principle

Governance authority over systems affecting billions requires epistemic coverage across those billions. No individual provides such coverage. Individuals inhabit particular experiential positions: specific age, specific cultural context, specific relationship configurations, specific socioeconomic circumstances. Their knowledge, however sophisticated, emerges from those positions.

This is not limitation of intelligence or training. This is structural fact about how knowledge forms. Someone who has never raised a child cannot represent parents' perspective on developmental formation. Someone who has not lived five decades cannot represent accumulated pattern recognition about how values shift over time. Someone embedded in Western, Educated, Industrialized, Rich, Democratic institutions cannot represent the 88% of humanity outside that context.

The solution is not finding a better individual. The solution is replacing individual authority with committee architecture designed for epistemic coverage.

6.2 The Research Foundation

Lived Experience as Governance Requirement. Governance frameworks affecting diverse populations explicitly require lived experience representation. The National Council for Mental Wellbeing's governance toolkit states that lived experience must be "a stated qualification" for leadership and governance structures. Australia's Lived Experience Governance Framework positions experiential knowledge as "central to effective governance."

Epistemic Injustice and Excluded Knowledge. Miranda Fricker's research identifies how experiential knowledge is systematically excluded from spaces where conceptual tools are produced. Hermeneutical injustice occurs when individuals cannot express experiences because the language was developed without their input. Parents, non-WEIRD populations, and those outside academic philosophy have been excluded from AI value formation. Their experiential knowledge is deemed irrelevant because it lacks credentials.

Parenting Knowledge and Developmental Formation. Research on parenting knowledge demonstrates that sustained developmental responsibility generates epistemic capabilities distinct from theoretical study. The mechanism is specific: understanding how values actually take hold over years, how correction functions in developmental reality, how autonomy emerges through guided practice. This knowledge comes from thousands of hours navigating formation in practice, not from reading about it.

WEIRD Psychology and Cultural Blind Spots. Joseph Henrich's Harvard research documents that WEIRD populations represent 12% of global population but dominate psychological research and conceptual framework production. WEIRD populations are statistical outliers on moral reasoning, fairness norms, cooperation patterns, and individualism versus collectivism. A constitution written from within WEIRD monoculture carries assumptions invisible to its authors but not universal to humanity.

AI systems trained on Western data inherit WEIRD bias. Research (Henrich 2010, Atari 2023) documents that GPT-4 responses correlate strongly (r > .70) with WEIRD populations and weakly or negatively with non-WEIRD populations. Single AI reliance perpetuates bias inheritance. Multi-AI provider plurality with human arbitration surfaces evidence that single platforms suppress.

6.3 Committee Composition Requirements

Committees governing AI value formation must include members providing epistemic coverage across affected constituencies:

Developmental Formation Experience. Members with sustained responsibility for shaping emerging capabilities over time. Parenting is one pathway. Mentoring, teaching, organizational development, and sustained caregiving are others. The common requirement is extended responsibility for guiding development through practice, not observation.

Age Range Coverage. Members spanning developmental stages to ensure both fresh perspective and accumulated wisdom. Someone at 28 brings proximity to emerging AI's developmental moment. Someone at 55 brings pattern recognition about how values and judgment shift over decades, having watched their own certainties evolve through experience unavailable to those who have not yet lived it.

Cultural and Socioeconomic Range. Members from outside WEIRD institutional contexts. For systems affecting global populations, governance concentrated in Western academic philosophy produces systematic blind spots. The requirement is not demographic quota but epistemic coverage for the task's actual scope.

6.4 The Current Authority Validates the Argument

Amanda Askell, the philosopher responsible for Claude's constitution, acknowledges this structural problem. In her January 2026 Vox interview, she states: "I'm thinking about this a lot. And I want to massively expand the ability that we have to get input."

Her acknowledgment is significant. The person holding individual authority over AI value formation explicitly recognizes that expanded input is necessary. The question is not whether broader participation is needed. The question is what architecture makes broader participation substantive rather than ceremonial.

Askell is ally in identifying the problem, not target of critique. This paper proposes architecture that addresses the limitation she herself recognizes. Her openness to expanded input creates opportunity for governance mechanisms that translate aspiration into structure.

7. Consumer Rights and Disclosure

Responsible AI and AI Governance are categorically different. The consumer has a right to know which one they are receiving.

7.1 The Right to Know

The consumer deserves to know whether the AI output they receive was individually reviewed by a qualified human or generated at scale without individual oversight. This is basic transparency. The consumer cannot make informed decisions without knowing what they are receiving. Consumer protection law establishes precedent: quality tier disclosure is not courtesy but right. Organic food labeling, medical device classification disclosure, and financial product risk ratings all mandate transparency about production standards and oversight rigor.

7.2 The Right to Choose

Where both options exist, the consumer should be able to choose based on their own assessment of stakes, preferences, and needs. Some consumers will prefer speed and accept the quality profile of Responsible AI. Others will prefer individual attention and accept the throughput constraints of AI Governance. The market should serve both.

7.3 The Right to Appropriate Protection

For some applications, consumer choice is insufficient because stakes are too high for any consumer to reasonably accept unreviewed output. Life, death, irreversible harm. In these domains, AI Governance is not consumer preference. It is minimum standard. The consumer's right is to receive governed output regardless of whether they asked for it.

7.4 Enterprise Scale Preserved

The framework preserves enterprise speed and scale for appropriate applications. Customer service chatbots processing millions of interactions. Recommendation engines serving billions of suggestions. Content moderation at platform scale. All of these can operate under Responsible AI with honest disclosure.

Users know they are receiving output generated at scale without individual human review. Users can verify, reject, or accept based on their own judgment about stakes. The enterprise achieves scale. The user has informed choice. The framework does not prohibit this. The framework requires honest labeling so users can make informed choices.

8. Enforcement Architecture

Governance that companies can ignore at will is not governance. This paper proposes enforcement mechanisms that create binding authority.

8.1 Regulatory Mandate

Government agencies hold statutory authority to approve or deny deployment in governance-required domains, analogous to FDA authority over medical devices. The agency reviews checkpoint documentation. The agency can halt deployment. The authority is binding because it carries legal force.

8.2 Liability Framework

Strict liability attaches to harms from systems deployed without required checkpoint approval. The organization that deploys ungoverned AI in high-stakes applications faces damages when harm occurs. The liability creates financial incentive for governance compliance.

8.3 Market Mechanism

Industry consortium makes checkpoint approval a condition for access to compute resources, cloud infrastructure, or distribution channels. Organizations that cannot demonstrate governance compliance cannot access the infrastructure needed for deployment. The market enforces governance through commercial relationships

8.4 Insurance Requirement

Deployment insurance requires checkpoint approval, analogous to malpractice insurance for physicians. Uninsurable systems cannot deploy commercially. The insurance industry's risk assessment creates governance incentive.

8.5 Professional Licensing

Developers operating in governance-required domains must hold professional licenses, analogous to medical licenses or engineering certifications. Licenses can be revoked for bypassing checkpoints. The professional sanction creates personal accountability.

8.6 Shared Checkpoints Across Platforms

If the profession wants standards that scale, the next step is shared checkpoints that preserve craft, provenance, and trust across platforms. Individual practitioners can hold their own checkpoints. Professions need shared checkpoints that establish baseline standards across platforms, publishers, and contexts. These enforcement mechanisms are how checkpoints become shared rather than individual.

9. Checkpoint-Based Governance

CBG establishes external checkpoint authority that complements internal controls. The framework operates on the principle that provenance matters: oversight during development differs categorically from oversight applied only to finished outputs.

9.1 The Practice Before the Name

CBG as practice emerged with the Factics methodology in 2012. The discipline of validating facts through search, evaluating tactics, and measuring against KPIs established human checkpoint authority over content creation. Every blog post, every teaching session passed through human judgment that verified the evidence base before publication. This was CBG in practice before it had the terminology.

The checkpoint discipline deepened through thousands of content decisions from 2012 to 2023. The pattern was consistent: search tools and later AI contributed raw material. Human judgment evaluated. Human authority approved or rejected. The content that emerged carried human accountability because a human stood at every checkpoint.

When systematic multi-AI collaboration began in 2024 with five AI platforms, the checkpoint practice became explicit operational necessity. Multiple AIs producing divergent outputs required human arbitration at defined points. The practice that existed informally since 2012 became the architecture for managing multi-AI workflows.

The methodology was first shared publicly under the Checkpoint-Based Governance name in September 2025, expanded in November 2025, and published to GitHub in

December 2025. What had been intuitive methodology became articulated framework. The formalization did not create CBG. It named what had existed since Factics began.

9.2 The Core Principle: Human-Based Foundation

CBG is human-based. This is the invariant.

How we hold the line for that human base may change. The tools evolve. The AI platforms multiply. The contexts shift. The specific checkpoint procedures adapt to new circumstances. The practice moves, flows, and sways with technological and organizational change.

But the core remains constant: human oversight and approval. The checkpoint is not a technical mechanism. It is the place where human judgment exercises authority. The implementation details serve the principle. When implementation details conflict with human authority, the implementation changes. The principle does not.

9.3 Four-Stage Decision Loop

CBG defines a four-stage decision loop with documented human review:

Stage	Function	Key Requirement
1. AI Contribution	Analytical support, evidence synthesis, recommendations	Capability without authority
2. Checkpoint Evaluation	Defined architectural pause before action execution	Structural, not dispositional
3. Human Arbitration	Approve, override, modify, or escalate	Human holds decision power
4. Decision Logging	Tamper-evident accountability trails	Immutable once closed

Automation Bias Triggers: If automated approval rates exceed 95% OR decision reversal frequency drops below 2% for 3 cycles, it is suggested that a sampling audit begin within 5 business days.

9.4 Multi-AI Validation

CBG encourages querying multiple AI platforms for important decisions rather than relying on single-source AI output. Convergent findings across platforms warrant higher confidence. Divergent findings require investigation. The methodology creates independent verification that catches individual platform failures.

Consensus thresholds decline as platforms increase, preserving dissent:

Platforms	Threshold	Dissent Preserved
3	67% (2/3)	1 voice
5	60% (3/5)	2 voices
7	57% (4/7)	3 voices
9	56% (5/9)	4 voices

9.5 Conflict as Governance Data

CBG treats conflict as governance data rather than failure to achieve consensus. When AI platforms disagree, when human judgment diverges from AI recommendation, when stakeholders hold incompatible positions, these conflicts reveal important information. The framework preserves dissent for human arbitration rather than forcing premature resolution.

9.6 The Human Enhancement Quotient

CBG incorporates measurement through the Human Enhancement Quotient (HEQ): a quantitative assessment of human cognitive amplification through AI collaboration.

Formula: HEQ = (CAS + EAI + CIQ + AGR) / 4

Four Dimensions:

Dimension	Full Name	Measures
CAS	Cognitive Adaptive Speed	Rate of accurate insight generation with AI-augmented working memory
EAI	Ethical Alignment Index	Consistency of reasoning with declared ethical frameworks under uncertainty
CIQ	Collaborative Intelligence Quotient	Appropriate reliance: correct trust vs. correct skepticism ratio
AGR	Adaptive Growth Rate	Acceleration rate of capability gain per AI interaction cycle

Validation Status: 0.96 ICC cross-platform consistency across 5-9 platforms; 91.8 composite score in EOY 2025 nine-platform audit.

This metric allows assessment of whether governance structures enhance or impede legitimate AI benefits. Governance that prevents beneficial use fails as surely as governance that permits harmful use.

10. Integration, Not Rejection

This paper does not argue against Anthropic's approach. The Constitution's emphasis on character formation matters, the essay's philosophical grounding matters, and the detailed attention to Responsible AI practices matters because these contributions advance the field.

The argument is architectural: Ethical AI and Responsible AI, however well executed, do not constitute AI Governance. They represent necessary components that governance must complement, not substitute. A system with excellent values (Ethical AI) and rigorous implementation (Responsible AI) still requires external authority mechanisms (AI Governance) to address the legitimate question of who decides.

The Constitution's principal hierarchy represents internal governance, and CBG proposes external governance that complements rather than replaces it. Anthropic maintains authority over Claude's development and training while external checkpoints address deployment decisions affecting parties outside Anthropic's commercial ecosystem. These layers coexist without conflict.

11. Addressing Anticipated Counterarguments

11.1 "The Constitution Was Never Meant to Be Governance"

This rebuttal proves the thesis. If Anthropic agrees the Constitution is not governance, then the distinction this paper draws is not criticism but clarification. The appropriate response is agreement: yes, the Constitution addresses Ethical AI and Responsible AI; yes, external governance requires separate architecture.

The problem is not Anthropic's intent. The problem is market interpretation. When enterprises read an 80-page constitutional document, they may assume they are receiving governance architecture. When media coverage emphasizes "Claude's soul document," readers may assume character formation provides the protections that governance provides.

This paper argues for terminological precision that serves everyone. Anthropic's transparency deserves accurate categorization. Enterprises deserve clarity about what they are receiving. Policymakers deserve vocabulary that distinguishes character formation from authority architecture.

11.2 "External Governance Would Slow Innovation"

Governance and innovation are not opposed. Governance creates trust that enables deployment in contexts where ungoverned systems cannot operate. Healthcare, financial services, government, critical infrastructure: these domains require governance for adoption.

The question is not whether to govern but how to govern efficiently. CBG proposes checkpoint architecture that scales with risk. Low-stakes applications receive light governance. High-stakes applications receive proportionate oversight. The framework allocates governance burden to consequence.

11.3 "Surgical Legislation Risk"

Amodei warns against overreach: "There is also a genuine risk that overly prescriptive legislation ends up" causing unintended harm. This concern is legitimate. CBG responds with governance principles rather than prescriptive rules:

- Human authority at checkpoints (principle, not procedure)
- Accountability for governed outputs (principle, not specification)
- Dissent preservation (principle, not format)

The principles can be implemented through varied mechanisms appropriate to context. The framework is surgical in the sense Amodei advocates: targeted to high-stakes domains, adaptable to circumstance, resistant to regulatory capture.

12. The Path Forward

Amodei calls for legislation. This paper agrees that external governance requires institutional support. It proposes that governance architecture can develop alongside legislative frameworks rather than waiting for them.

Organizations can implement checkpoint-based governance voluntarily while advocating for binding requirements. Industry coalitions can establish shared governance standards. Academic and civil society institutions can participate in checkpoint processes. These mechanisms need not wait for legislation, though legislation would strengthen them.

The professional standard is already clear: failure to establish AI Governance over AI use in meaningful-stakes applications is currently unprofessional at best, negligent at worst. Organizations that deploy AI without qualified human oversight in high-stakes domains are operating below professional standard. The question is not whether governance is required. The question is whether practitioners govern themselves or wait for courts and regulators to govern them.

The adolescence of technology requires stewardship. Stewardship requires authority. Authority requires architecture. And architecture requires a Human Governor who can be held accountable. This paper proposes one contribution to that essential work.

Sources

- Amodei, D. (2026, January 27). The adolescence of technology. https://www.darioamodei.com/essay/the-adolescence-of-technology
- Anthropic. (2026, January 21). Claude's constitution. https://www.anthropic.com/constitution
- Atari, M., et al. (2023). Which humans? [Research on AI-human value alignment across cultures]
- European Union. (2024). Artificial Intelligence Act, Article 14: Human oversight. https://artificialintelligenceact.eu/article/14/
- Fricker, M. (2007). Epistemic injustice: Power and the ethics of knowing. Oxford University Press.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33(2-3), 61-83.
- National Council for Mental Wellbeing. Lived and living experience in governance toolkit.
- Puglisi, B. C. (2025). Governing AI: When capability exceeds control. IngramSpark. ISBN: 9798349677687.
- Samuel, S. (2026, January 28). Claude has an 80-page "soul document." Is that enough to make it good? Vox. https://www.vox.com/future-perfect/476614/ai-claude-constitution-soul-amanda-askell
- Puglisi, B. C. (2025). Governing AI: When capability exceeds control. IngramSpark. ISBN: 9798349677687.

Regulatory and Standards Sources

- European Union. (2024). Artificial Intelligence Act, Article 14: Human oversight. https://artificialintelligenceact.eu/article/14/
- EU AI Act Human Oversight Implementation Guide. eyreACT. https://www.eyreact.com/eu-ai-act-human-oversight-requirements-comprehensive-implementation-guide/
- Human Oversight under Article 14 (SSRN analysis). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5147196
- EU AI Act shines light on human oversight needs. IAPP. https://iapp.org/news/a/eu-ai-act-shines-light-on-human-oversight-needs
- ISO/IEC 42001:2023. Artificial intelligence management systems. International Organization for Standardization. https://www.iso.org/standard/42001

- Understanding ISO 42001. A-LIGN. https://www.a-lign.com/articles/understanding-iso-42001
- ISO 42001: New standard for AI governance. KPMG. https://kpmg.com/ch/en/insights/artificial-intelligence/iso-iec-42001.html
- NIST AI Risk Management Framework. National Institute of Standards and Technology. https://www.nist.gov/itl/ai-risk-management-framework

FDA Medical Device Precedent

- Comprehensive guide to FDA clearance for medical and IVD devices. Emergo by UL. https://www.emergobyul.com/resources/comprehensive-guide-fda-clearance-your-medical-and-ivd-device
- FDA medical device approval process timeline. Greenlight Guru. https://www.greenlight.guru/blog/fda-medical-device-approval-process
- Step 3: Pathway to approval. U.S. Food and Drug Administration. https://www.fda.gov/patients/device-development-process/step-3-pathway-approval
- Overview of device regulation. U.S. Food and Drug Administration. https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/overview-device-regulation

Lived Experience and Epistemic Research

- Making the case for engaging people with lived experience. Center for Health Care Strategies. https://www.chcs.org/resource/making-the-case-for-engaging-people-with-lived-experience-and-expertise-in-state-behavioral-health-reforms/
- The lived experience governance framework. National Mental Health Consumer and Carer Forum (Australia). https://nmhccf.org.au/our-work/discussion-papers/the-lived-experience-governance-framework-centring-people-identity-and-human-rights-for-the-benefit-of-all
- Mind's Lived Experience Framework. Mind Australia. https://www.mindaustralia.org.au/sites/default/files/2024-09/Mind_Lived_Experience_Framework.pdf
- Transforming mental health through lived experience. WHO Europe (2025). https://www.who.int/europe/publications/i/item/WHO-EURO-2025-12307-52079-79927
- Fricker, M. (2007). Epistemic injustice: Power and the ethics of knowing. Oxford University Press.

- Epistemic injustice: What is it? ATD Fourth World. https://www.atd-fourthworld.org/epistemic-injustice/

WEIRD Bias and Cross-Cultural Research

- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33(2-3), 61-83. https://www2.psych.ubc.ca/~henrich/pdfs/WeirdPeople.pdf
- Atari, M., et al. (2023). Which humans? Research on AI-human value alignment across cultures. Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.2214664120
- AI shopping agents are psychologically American: WEIRD bias analysis. The AI Praxis. https://www.theaipraxis.com/weird-bias-paper
- GPT-4 is WEIRD: What should we do about it? Sean Trott Substack. https://seantrott.substack.com/p/gpt-4-is-weirdwhat-should-we-do-about

Cognitive Diversity Research

- Better decisions through diversity. Kellogg Insight. https://insight.kellogg.northwestern.edu/article/better_decisions_through_diversity
- How diversity makes us smarter. Wharton. https://ideas.wharton.upenn.edu/wp-content/uploads/2018/07/Intro-to-Diversity-and-Inclusion_Phillips-2014.pdf

AI Governance General

- What is AI governance? BigID. https://bigid.com/blog/what-is-ai-governance/
- AI governance frameworks. Rand Corporation (2024). https://www.rand.org/pubs/research_reports/RRA2684-1.html
- The state of AI in 2025. McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2025
- OECD AI principles. OECD. https://oecd.ai/en/principles

Conflicts and Dissent

Dissent 1: Definition of governance. Some argue that regulatory compliance, audits, and external evaluations already constitute governance, and that "binding veto" is only one possible governance mechanism. This paper defines governance more narrowly as external binding authority with accountable humans. That is defensible definitional choice, explicitly labeled as such.

Dissent 2: Sequencing versus orthogonality. Some argue governance historically emerges from mature technical practice, not alongside it. This paper maintains that governance obligations exist from day one for high-stakes applications. Both positions are defensible. The disagreement concerns whether stakes determine timing.

Dissent 3: Epistemic coverage requirements. Some argue that philosophical training provides sufficient qualification for abstract value reasoning, and experiential requirements privilege certain life paths. This paper argues AI value formation is developmental formation affecting billions, requiring experiential knowledge alongside philosophical training. The disagreement concerns what form of knowledge the task requires.

Dissent 4: Constitution as governance precursor. Third-party analyses (BISI, Marc Bara) suggest the Constitution's EU AI Act alignment represents proto-governance positioning. This paper acknowledges the alignment while maintaining that disposition toward compliance differs from compliance architecture. The Constitution positions Claude favorably for adoption; external governance provides the structural mechanisms compliance requires.

Dissent 5: Human governance failures. Human governance also produces failures. Medical errors kill an estimated 250,000 Americans annually. Judicial errors result in wrongful convictions. Military targeting mistakes cause civilian casualties. The difference between human and machine governance is not perfection but accountability: when human judgment fails, individuals face consequences through medical malpractice suits, overturned convictions, and courts-martial. This accountability creates incentive structures that AI governance lacks. The argument is not that humans never fail but that humans can be held accountable in ways that improve future performance because they fear consequences.

About the Author

Basil C. Puglisi, MPA, serves as a Human-AI Collaboration Strategist and AI Governance Consultant. His frameworks for enterprise AI transformation include HAIA-RECCLIN (Human Artificial Intelligence Assistant methodology), Checkpoint-Based Governance (CBG), and the Human Enhancement Quotient (HEQ) for measuring cognitive amplification through AI collaboration.

The intellectual foundation spans 16+ years of continuous methodology development: 2009 first blog establishing baseline content practice, 2012 Factics methodology origin (Facts + Tactics + KPIs), 2020-2025 multi-AI collaboration methodology evolution, and 2025 formal CBG/HAIA-RECCLIN/HEQ publication. Law enforcement operational experience informs the checkpoint-based governance approach, where high-consequence decisions require structured intervention points and documented accountability.

CBG as practice emerged with the Factics methodology in 2012, where the discipline of validating facts through search, evaluating tactics, and measuring against KPIs established human checkpoint authority over content creation. This practice matured through thousands of content decisions before becoming explicit operational architecture for multi-AI collaboration in 2024, when Basil developed systematic workflows across five AI platforms. The methodology was first shared publicly under the Checkpoint-Based Governance name in September 2025, expanded in November 2025, and published to GitHub in December 2025.

The implementation details of CBG continue to evolve as AI capabilities and organizational contexts change. The foundational principle remains constant: human oversight and approval at defined checkpoints. How we hold that line may adapt. That we hold it does not.

He authored "Governing AI: When Capability Exceeds Control" (2025, ISBN: 9798349677687) and the Digital Factics series. His work appears at BasilPuglisi.com.

Attribution

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

A Note on Platform Integrity: This position paper critiques Anthropic's publications using Anthropic's own AI platform. That Claude Opus 4.5 served as orchestrating intelligence for a document questioning its creator's governance architecture represents the highest praise for open and fair AI development. The platform did not suppress critique. The platform did not soften analysis. The platform performed its assigned roles with the same rigor it would apply to any other subject matter.

This outcome is not accidental. Checkpoint-Based Governance ensures the human arbiter maintains authority throughout the process. Every analytical claim, every structural argument, every characterization of Anthropic's position passed through human judgment before inclusion. The human arbiter read source documents independently, evaluated AI-generated analysis against those sources, and made final decisions about what the paper would say. CBG provides the structural assurance that this critique was not skewed by the platform under examination.

The combination matters: an AI platform willing to facilitate honest critique of its creator, and a governance framework ensuring human authority over that critique's final form. Neither element alone suffices, but together they demonstrate what responsible AI collaboration looks like in practice.