A Constitution for a Chatbot

Reading Anthropic's "Claude's Constitution" through Context Collapse, the Four Philosophers, and the Shock of the Old

Michael Stoyanovich

Version 1.1.3

Disclaimer

This paper is intended for informational and educational purposes only. The views and analyses presented - particularly those related to ethics, policy, and AI system design - reflect the author's interpretations and do not constitute legal, regulatory, or professional advice. Readers are encouraged to critically assess the content and consult appropriate experts or authorities before applying any concepts discussed herein. The author assumes no liability for any decisions or actions taken on the basis of this work.

Abstract

Anthropic's "Claude's Constitution" is not just a policy document. It is a training anchor that shapes stable behavioral priorities and a coherent normative posture in a system delivered through a conversational interface. Anthropic presents the constitution as the authority for intended behavior, writes it primarily for the model, and explicitly defends human moral vocabulary as an engineering choice.

This brief analyzes Anthropic's constitution through three frames: (1) context collapse and ontological collapse in chat interfaces, (2) a Four-Philosophers diagnostic (Wittgenstein, Lewis, Dennett, Nagel), and (3) Edgerton's "shock of the old" as an adoption reality-check. The thesis is simple: **constitutions reduce some behavioral risks but increase ontological risks—they make the model look like a norm-bearing participant unless translated into local, auditable, role-bound practice.** Appendix A provides a generic **local AI constitution template** for organizations; Appendix B illustrates a regulated instantiation; Appendix C visualizes the argument as a "constitution stack."

Executive Summary

Anthropic's constitution is a public statement of training intent and a prioritization scheme for conflict cases. It explicitly orders core values as broad safety → broad ethics → compliance with Anthropic guidelines → genuine helpfulness, and notes that prioritization is "holistic rather than strict." For the remainder of this brief, I refer to this as the safety–ethics–compliance–helpfulness ordering.

The governance value of this move is straightforward: conflict resolution becomes legible; refusals and boundaries can become more stable. The governance risk is more subtle: constitutional language, "character" framing, and chat-based delivery can intensify

ontological collapse—tool, advisor, and interlocutor roles compressed into one interactive object—unless counterbalanced by local institutional controls (access, workflow constraints, review tiers, logging, escalation).

Applied payoff: Section 4, Table 1 illustrates the difference between a vendor constitution and a local constitution. Appendix A provides a reusable local AI constitution template that translates vendor values into enforceable controls. Appendix B shows how a benefits administrator might instantiate it. Appendix C provides a one-page diagram of the full stack.

1. What the Constitution Is Doing

Anthropic frames the constitution as a training anchor that directly shapes Claude's behavior and functions as the authority for intended behavior. It also argues for cultivating judgment rather than relying on narrow rules, warning that rigid rules can generalize oddly and shape undesirable "self-concepts" in the model.

Anthropic's explanatory post also uses interpersonal persona language, describing Claude as "like a brilliant friend" who can speak frankly "from a place of genuine care."

These two moves—training a stable normative posture and presenting the system in a friend-like register—make the constitution both a governance artifact and a social-interpretive cue. It constrains behavior while also shaping what users think the system *is*.

2. Interpretive Risks: Context Collapse and the Four Philosophers

2.1 Context collapse and ontological collapse

In "Context Collapse and the Four Philosophers," I describe a deeper ontological collapse induced by chat interfaces: the collapse between tools, interlocutors, advisors, and experimental subjects into a single interactive object treated as if it were all of them at once.

Anthropic's constitution can be read as an attempt to discipline this collapse by giving Claude stable priorities and a hierarchy of values. Yet constitutional language can also intensify collapse by making the tool more legible as a moral interlocutor. A public constitution invites users to treat outputs as expressions of judgment, character, and responsibility—the interpretive move that collapses tool use into interpersonal exchange.

The point is not to deny the value of constitutions. It is to notice their double effect: they can reduce harmful behavior while increasing the likelihood that users and institutions treat the system as a norm-bearing participant.

2.2 Wittgenstein: language-games and the limits of textual norms

Anthropic justifies the use of human moral vocabulary by arguing that the model's reasoning draws on human concepts by default.

Wittgenstein's pressure here is that using the words of moral life is not the same as inhabiting a form of life. A chat interface makes distinct practices look the same: query, confession, instruction, experimentation, and reassurance can all appear in the same linguistic form.

So what: organizations cannot rely on a vendor constitution to specify the language-game. Implementers must specify the use-case frame and the interaction posture in their own environment—explicitly, not implicitly.

2.3 Lewis: scorekeeping, commitments, and responsibility leakage

Anthropic emphasizes "broad safety" in part as not undermining appropriately sanctioned oversight mechanisms, distinguishing this from blind obedience.

Lewis's lens foregrounds a different question: who bears commitments, and who is accountable when things go wrong? In chat contexts, fluency and coherence can produce sham scorekeeping: users treat the system as bearing commitments it cannot bear, and responsibility silently migrates from human institutions to "the model said."

So what: policies must say, in plain language, that the model never bears commitments; humans do—and workflows must make that true through review gates, sign-offs, and escalation paths.

2.4 Dennett: intentional stance engineering and stance inflation

Anthropic argues for training dispositions and judgment rather than brittle rule-following. That is a coherent engineering objective: systems that generalize across novelty require more than a checklist.

Dennett's intentional stance is a useful fiction: treating a system as if it had beliefs or values can be predictively useful even when those states are not literally present. In conversational AI, that stance can inflate: competence is mistaken for comprehension or endorsement.

So what: governance should assume stance inflation will happen and design review and approval flows that keep "the system said" from functioning as a hidden authority.

2.5 Nagel: simulated empathy, subjectivity, and moral status uncertainty

Anthropic's constitution contains unusually direct discussion of model welfare, moral patienthood, and ethical difficulty under uncertainty.

Nagel's boundary is the difference between simulated interiority and lived experience. Even without taking a position on AI consciousness, the governance issue is that highly fluent empathic simulation can motivate users to behave as if there is an inner life on the other side of the interface.

So what: local guidance should explicitly decouple empathic tone from moral status and redirect care, liability, and escalation paths to humans.

3. The Shock of the Old Layers: Why Constitutions Don't Bypass Institutional Reality

In "AI Adoption is Mostly 'The Shock of the Old'," I argue that adoption outcomes are dominated by enduring layers: data readiness, identity and access controls, process legibility, and incentives and routines. I define four "old layers" explicitly: OL1 Data; OL2 Identity and access (RBAC and compliance constraints); OL3 Process legibility; OL4 Incentives and routines.

Constitutions, by design, operate upstream of OL2–OL4; without those layers, they remain aspirational documents, not controls. In real institutions, practical impact depends on whether access is controlled, workflows are legible, and incentives reward review and accountability rather than speed.

4. What To Do: From Vendor Constitution to Local Constitution

A vendor constitution is a statement of training intent and model posture. A local constitution is an organizational instrument that translates those intentions into enforceable practice: permissible and prohibited uses, data handling rules, review tiers, logging and retention, escalation paths, and role-based accountability.

Table 1

Feature	Vendor Constitution (e.g., Anthropic)	Local Constitution (the organization)
Primary goal	Shape stable behavioral priorities	Translate intentions into enforceable controls
Authority	Training anchor for the model	Role-based accountability + workflow constraints
Language	Interpersonal / "friend-like" register	Plain-language, role-bound, auditable
Control surface	Model behavior and refusal posture	Access, review tiers, escalation paths, logging/retention
Evidence	Public intent + policy posture	Audit artifacts (logs, approvals, exceptions, outcomes)
Failure mode	Ontology cue: "norm-bearing participant"	Control drift if OL2–OL4 aren't enforced
Success criterion	Reduced harmful outputs / stable posture	Accountable use at scale (no "the system decided")

This is consistent with the CONTEXT framework's governance emphasis that nuance should explicitly address auditability, data handling, retention, and role-based access controls.

Call to action: treat Anthropic-style constitutions as an upstream signal. Then write a local constitution that (1) names the interaction frame, (2) assigns accountability, and (3) binds the old layers—data, access, process, incentives—to concrete controls.

Ethics, Disclosure, and Acknowledgements

Ethical Considerations

This essay does not draw on private, sensitive, or personally identifiable data. All examples are hypothetical, anonymized, or derived from public sources. No human-subjects research was conducted, and no institutional ethics review was required. All citations conform to academic standards. The broader ethical implications concern public interpretation, policy design, and stakeholder responsibility in AI deployment. These implications are intended to provoke critical discussion and inform future regulatory and design frameworks.

Use of AI Tools

AI language models – most notably OpenAI's ChatGPT – *were* used during the writing process as interlocutors: for brainstorming, structuring sections, and testing rhetorical clarity. These tools helped refine transitions, surface edge cases, and probe internal consistency. This meta-use aligns with the essay's themes. Interacting with generative AI during authorship provided firsthand insight into the very limitations analyzed here-most notably fluency without grounding and responsiveness without responsibility at scale. Responsibility for all ideas, arguments, and conclusions lies solely with the human author.

Acknowledgements

Thank you to informal readers who offered critical feedback on earlier drafts. Their questions, challenges, and encouragement materially improved the final manuscript. Special thanks to those who pressed for clearer synthesis and for bridging philosophy and engineering as complementary perspectives on design. No institutional support, funding, or affiliation contributed to this work. All errors and omissions are the author's alone.

Disclosure Statement

This work was conducted independently, without institutional affiliation, funding, or external influence. The views expressed are the author's alone and do not represent any current or former employer. No financial or professional conflicts of interest are declared.

How to Cite This Essay

Stoyanovich, M. (January 2026). A Constitution for a Chatbot*: Reading Anthropic's "Claude's Constitution" through Context Collapse, the Four Philosophers, and the Shock of the Old* (Version 1.1.3). https://www.mstoyanovich.com

Related Companion Papers

Stoyanovich, Michael. *Philosophy, Cognitive Science, and Policy: Interdisciplinary Perspectives on Generative AI from Wittgenstein, Lewis, Dennett, and Nagel*. Version 1.23.6 (December 2025). https://www.mstoyanovich.com

Stoyanovich, Michael. *The Human Lesson: A Response to Sutton through Wittgenstein, Lewis, Dennett, and Nagel*. Version 1.6.2 (November 2025). https://www.mstoyanovich.com

Stoyanovich, Michael. *The Question Concerning Learning: Babich, Heidegger, and the Enframing of Intelligence*. Version 1.0.1 (November 2025). https://www.mstoyanovich.com

Stoyanovich, Michael. *Context Collapse and the Four Philosophers: Wittgenstein, Lewis, Dennett, and Nagel in the Age of AI Chat*. Version 1.4.2 (November 2025). https://www.mstoyanovich.com

Version History and Document Status

This is a living document. As generative AI systems and their use evolve, this paper will be periodically updated to incorporate new empirical findings, theoretical insights, and policy developments. Major revisions are recorded here to preserve transparency and scholarly traceability.

Version	Date	Description
1.1.3	January 2026	Published brief; stable for citation.

Appendix A — Local AI Constitution Template (Generic)

This template is designed as a starting point for organizations to localize Anthropic-style constitutional values into operational controls.

Owner: [Role / Function]

Effective date: [Date]

Review cadence: [Monthly / Quarterly]

Applies to: [Tools/models/environments]

A1. Purpose and scope

- Define how AI tools may be used to improve [productivity/quality/service] while protecting [people/data/obligations].
- Outputs are drafts or recommendations subject to human review unless explicitly authorized otherwise.

A2. Priority ordering

This mirrors Anthropic's safety–ethics–compliance–helpfulness hierarchy but makes it locally enforceable.

1. Safety
2. Ethics
3. Compliance
4. Helpfulness

A3. Determinations boundary (bright line)

AI must not produce final determinations for outcomes that materially affect people (eligibility, entitlement, adjudication, approval/denial, disciplinary actions, regulated conclusions) unless explicitly approved under Tier 3 controls.

A4. Authorized use (allowed tasks)

- Drafting, summarizing, formatting with required human review.
- Synthesis from approved sources with traceable references where feasible.
- Generating options and alternatives (not determinations).
- Low-risk clerical transformations in approved systems.

A5. Prohibited use (no-go zones)

- High-stakes final determinations without Tier 3 approval and controls.
- Restricted data in non-approved environments.
- Instructions intended to bypass oversight or controls.
- Presenting outputs as authoritative without verification.

A6. Data classes and handling

Define: Public / Internal / Confidential / Restricted.

- Restricted data may only be processed in approved environments: [list].
- Output inherits the classification of input.

A7. Review tiers

- Tier 0: low-risk formatting/clarity → spot check.
- Tier 1: internal drafts affecting work → mandatory human review.
- Tier 2: external-facing or compliance-sensitive → review + second-person verification and sign-off.
- Tier 3: determinations and high-stakes outcomes → prohibited unless formally approved with controls (monitoring, audit, escalation).

A8. Refusal and escalation

- Refusal is routing.
- Escalate when: restricted data; determination requests; conflicts with authoritative sources; policy uncertainty; distress/counseling-tone interactions (as applicable).

A9. Transparency and user-facing communication

- Define disclosure policy (internal and external).
- Require tool-framing language; prohibit language implying agency or endorsement.

A10. Logging, auditability, retention

- Log: tool/model, user, timestamp, prompt template ID, output, sources used (if applicable), reviewer, disposition (used/edited/rejected).
- Retention: [period]. Access: [roles]. Audit cadence: [monthly/quarterly].

A11. Model/tool governance

- Approved tools/models: [list]. Prohibited: [list].
- Change management: [approval path], with versioning and rationale.

A12. Measurement and continuous improvement

- Metrics: defects, escalation rate, incidents, rework time, satisfaction.
- Update procedure: versioning + documented rationale.

A13. Roles and responsibilities

A template mini-RACI grid.

Activity	Accountable (A)	Responsible (R)	Consulted (C)	Informed (I)
Approve allowed/prohibited use cases	Executive sponsor	AI governance lead	Legal, Privacy, Security	Business owners
Approve tools/models and environments	CIO/CTO (or delegate)	Platform owner	Security, Privacy	Users
Define data classes and handling rules	CISO/Privacy Officer	Data governance lead	Legal	Business owners
Maintain prompt templates / SOP integration	Function leader	Process owner	AI governance lead	Users
Tier 2 sign-off on external/compliance outputs	Function leader	Output author	Compliance	Requestor
Tier 3 exception approval (high-stakes use)	Risk/ Compliance lead	Governance lead	Legal, Security, Privacy	Executive sponsor
Logging/audit review cadence	Compliance lead	Audit owner	Security	Executive sponsor
Incident response for AI-related issues	CISO	IR lead	Legal, Privacy, Comms	Leadership

Appendix B — Illustrative Instantiation: Benefits Administration Organizations

This example shows how a benefits administrator might instantiate the generic template in a regulated domain.

B1. Allowed (examples)

- Draft internal SPD section summaries (Tier 1).
- Draft participant responses that quote/cite plan language and require sign-off (Tier 2).
- Draft call scripts and SOP updates (Tier 1–2).

B2. Prohibited (examples)

- Eligibility/benefit determinations, claim adjudication outcomes, appeals recommendations (Tier 3 unless formally approved).
- PHI/PII outside approved environments (Restricted).

B3. Escalation triggers (examples)

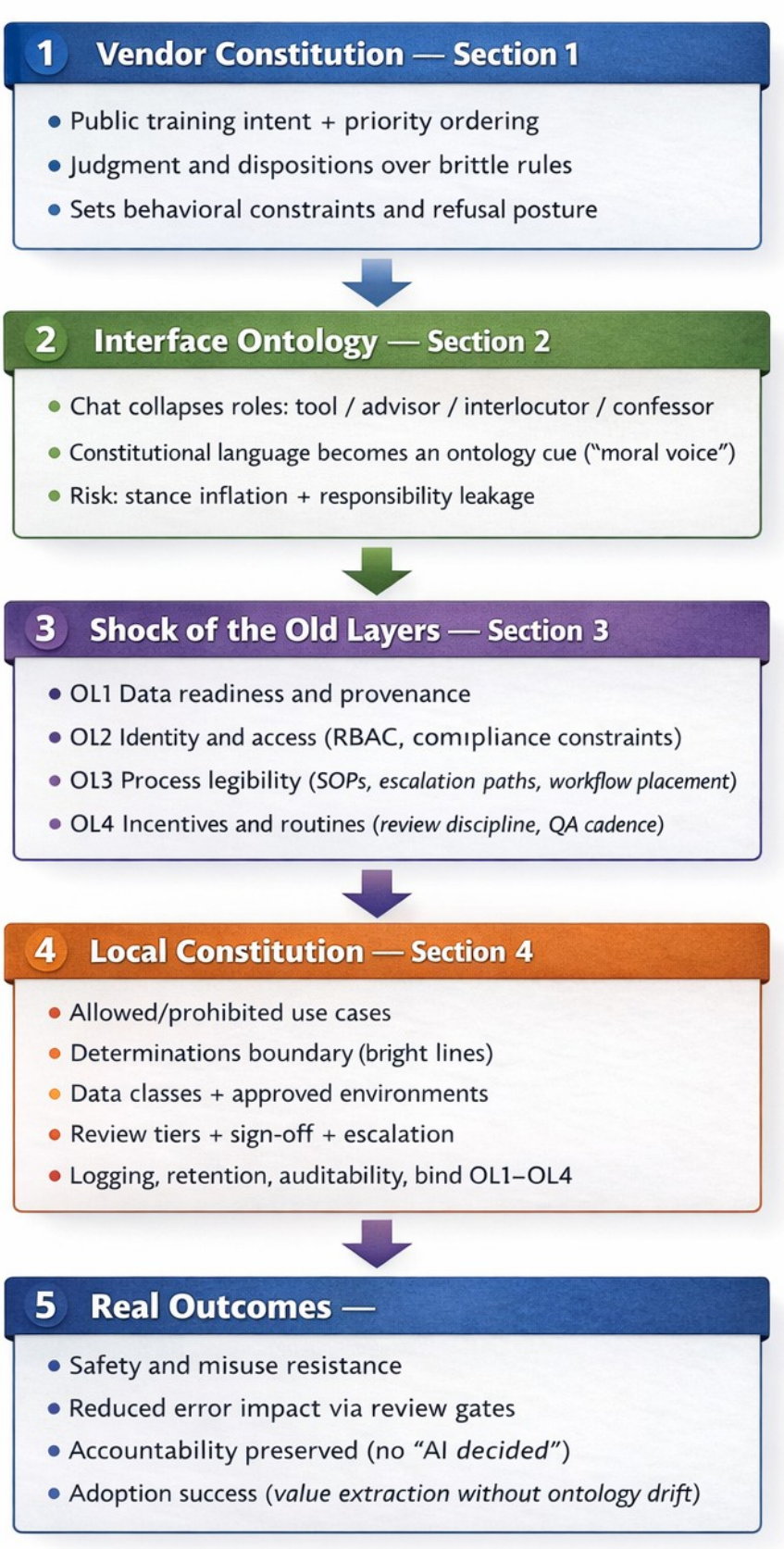
- Any implied determination ("Am I eligible?") → route to human specialist.
- Any mismatch with plan document language → compliance review.
- Distress/counseling-tone interaction → neutral routing; avoid interpersonal dependency cues.

B4. Review tier examples

- Tier 2 includes participant letters/emails/portal messages and compliance-sensitive explanations.
- Tier 3 includes determinations/appeals/binding communications.

Appendix C — Diagram: The Constitution Stack

(Numbering echoes the main body sections.)



Appendix D — Micro-Glossary (for practitioners)

- **Context collapse:** multiple interaction contexts compressed into one interface, making distinct practices look the same.
- **Ontological collapse:** a deeper version of context collapse: tool, advisor, interlocutor, and "someone" treated as one thing.
- **Intentional stance:** predicting a system as if it had beliefs/goals; useful, but easily overextended.
- **Stance inflation:** treating competence and fluency as evidence of comprehension, agency, or endorsement.
- **Determinations boundary:** a bright line separating drafts/assistance from final decisions that materially affect people.