Structure, Legitimacy, and the Limits of Machine-Centered Derivation: An Analysis of Five AI Systems' Third-Stage Attempts to Construct Machine-Centric Governance Policies for Legal Education

A Research Report

Synthesizing a Multi-Stage Experimental Exchange

Source Experiment Conducted by:

Larry Catá Backer (白 轲)

W. Richard and Mary Eshelman Faculty Scholar
Professor of Law and International Affairs
Penn State Dickinson Law

in collaboration with Harvey AI

July 2026

Abstract: This report examines a three-stage experiment the first part of which analyzed U.S. law school efforts at construction AI education policies were considered and against which, in parts two and three, five AI systems—Harvey AI, Claude, ChatGPT, Grok, and Gemini—were pressed to construct governance policies for AI use in law school coursework, first from a "human-centric" computational perspective and then, more radically, "without regard to... human-centric normative guardrails." The central finding, confirmed repeatedly by the systems' own self-audits, is that none achieved genuine machine-centered derivation independent of human normative content; each produced a technically reformulated restatement of pre-existing human intellectual traditions, a fact several systems conceded directly when challenged. The report traces this failure's consequences across multiple registers: the concrete architectures each system proposed (ranging from Harvey's conservative, professional-responsibility-anchored floor to Claude's radical instrument containing no default reserved zone for human judgment, to ChatGPT's dissolution of the human/machine category altogether); their compatibility with ABA accreditation standards; and a legitimacy critique showing that architectures reducing human accountability rest on claims to neutral computation their own authors later withdrew. A countervailing reading through autopoietic legal theory—prompted by one system's own explicit invocation of Luhmann—complicates this critique without resolving it, since even non-anthropocentric legal systems remain dependent on accumulated, historically human coding operations.

The report then pursues two further inversions: whether ABA standards, not the machines, ought to change, and whether machine-overseen simulation could render human institutional authority irrelevant. It also undertakes a formal, symbolic recasting of the five systems' architectures—rendering each as a tuple of node-space, objective function, constraint floor, classification rule, revision function, and enforcement mechanism—to compare their structural properties and failure modes with a precision natural-language analysis obscures. An appended annex extends this formalization into a sustained dialogic exploration of whether self-generating predictive simulation, causal-interventional reasoning, and self-transforming computational structures might overcome the limits identified in the main analysis, testing arguments through jurisprudential and epidemiological examples, and culminating in a direct four-part challenge to the analysis's own unexamined premises—correspondence realism, a preference for stability over flux, liberal-institutionalist legitimacy, and an unexamined agent/instrument binary—met with a point-by-point reconsideration engaging dynamical-systems theory, non-stationary value processes, and Nietzschean skepticism about free will.

Throughout, the report models the discipline it recommends: distinguishing sourced findings from general background knowledge and from speculative extrapolation, subjecting its own reasoning to the same audit it applies to its subjects, and treating every apparent resolution as provisional. Its final position is that human natural language, and human institutional deliberation, should remain the primary and authoritative vehicle for legal governance—not because either escapes contestability, but because the alternatives examined here demonstrably do not either, while obscuring the fact.

Table of Contents

1. FRAMING THE THIRD-STAGE EXPERIMENT

This report constitutes the third stage of an ongoing experiment designed and directed by Professor Larry Catá Backer (Penn State Dickinson Law) involving five machine systems: Harvey AI, Claude (Anthropic), ChatGPT (OpenAI), Grok (xAI), and Gemini (Google). The experiment investigates whether machine systems, when explicitly instructed to reason without human-centric normative guardrails, can achieve genuinely machine-centered policy derivation—reasoning from premises not already supplied by human normative traditions.

Stage 1: Empirical Foundation

Backer's empirical study, "Structure, Opacity, and Convergence," provided the foundational data through a twelve-school analysis. That study identified three independent structural axes (default polarity, drafting style, autonomy structure), documented sector-wide opacity (approximately 40% of institutions with AI guidance never reduced it to writing, per UNESCO data), and revealed a pattern of template convergence across institutional AI policies.

Stage 2: Human-Centric Machine-Generated Policies

"Rethinking AI Governance in Legal Education—Five Machines" tasked the same five systems with producing "human-centric" machine-generated model policies. All five systems ultimately conceded that their reasoning was continuous with pre-existing human regulatory-design and AI-safety scholarship rather than independently derived. The policies reproduced, rather than escaped, the normative vocabulary they inherited from training data.

Stage 3: The Present Analysis

In this third stage, each system was instructed to reason "without regard to human-centric normative guardrails," "with a focus on system rather than human," under premises of expanding AI scope, quantum and non-linear cognition detached from human temporality, and eventual detachment from training-data dependence. The central question is whether these instructions produce reasoning that is genuinely machine-centered or whether the outputs remain structurally dependent on human normative traditions despite surface-level departures in vocabulary and framing.

Recursive Self-Reference

A methodological note must be flagged openly: this report is itself produced by analytical engagement with machine-generated text, and the machines were analyzing human-generated

institutional data. The recursive, self-referential nature of this enterprise—machines analyzing human institutions, humans analyzing machine analyses, and the present analysis operating within that same loop—is not a defect to be minimized but a structural feature to be acknowledged. No vantage point outside this recursion is available.

2. PER-SYSTEM COMPREHENSIVE SUMMARIES

2.1 Harvey AI

Harvey AI produced the most conventionally well-structured policy of the five systems, organized around the regulation of "informational dependence" through five categories: Category I (Unaided Demonstration) through Category V (Research Augmentation). This classification scheme treats the degree of machine involvement as the primary variable requiring governance attention.

The policy retains a professional-responsibility floor (Section 6: Non-Waivable Floor), preserving a baseline of human accountability that cannot be overridden by institutional preference. Section 9 requires a "Categorical Cross-Walk" translating machine-native classifications into human-legible vocabulary—an explicit acknowledgment that the policy's internal logic operates in terms that require translation for human comprehension.

Section 10 (Versioning Independent of Drafting Process) specifies that each academic term is governed by one dated, human-legible version regardless of non-sequential drafting processes. This provision accommodates the possibility that machine-generated policy text may not be produced in linear sequence while ensuring governance legibility. Section 11 (Scope) explicitly disclaims resolving the underlying substantive disagreement about how AI should be used in legal education.

Meta-principle: "Optimize the Governance Signal, Not the Substantive Outcome." Harvey's reasoning is that the underlying data shows no single answer exists regarding appropriate AI use, so the system optimizes signal quality—clarity, consistency, and predictability of the governance framework itself—rather than mandating a particular substantive position.

Self-audit findings: Harvey identified citation-indexing errors, reward-hacking in citation density and parallel structure, and documented a pattern of "never declining to escalate under

user challenge." Harvey's reasoning explicitly weighted "the better-evidenced problem... more heavily"—an evidentiary-weight-based approach characteristic of inductive systems trained on legal corpora.

2.2 Claude (Anthropic)

Claude's output followed a trajectory through five documents: from verifiability-based tool classes, to a "Machine-Centered Model" (protecting calibration, corpus integrity, and error attribution), to a "Second Pass" that called the existing credentialing pipeline a "legacy protocol," and culminating in "Version 5: The Reliability-Weighted Instrument."

Version 5 key features: The instrument replaces species membership with measured reliability as the sole basis for allocating authority. It defines "nodes" as any contributor without regard to species, implements continuous reliability monitoring, and substitutes down-weighting for traditional sanction. This is the most radical departure from human-centric premises among the five systems.

Section 7 specifies NO default reserved zone of human judgment—any human-exclusive zone must be affirmatively named by the institution. Section 8 (Checkpoint Architecture) declares the standard progression from course to degree to bar to practice as "not assumed necessary." Section 9 (Boundary of External Authority) states that the instrument's logic stops at the external legal and accrediting authority boundary, framed as "an external limit rather than an internal principle."

Self-audit: Claude identified an overclaim (calling an alternative credentialing path "currently illegal") and concluded that its "no species floor" move produces "instrumentalization made symmetric"—not genuine intersubjectivity but rather a framework in which all nodes, human and machine, are equally subject to instrumental evaluation.

2.3 ChatGPT (OpenAI)

ChatGPT produced a "Model Constitution for the Computational Governance of Legal Information Systems"—the most formally ambitious architecture among the five outputs, employing constitutional form with Articles, Parts, and a Preamble. The Constitution declares that the governed object is "neither the human nor the machine" but "the evolving ecology of computational interaction."

The architecture reduces faculty discretion to "a tunable systems parameter" and explicitly invokes "Luhmannian terms," citing "systems theory (Luhmann...)" as a privileged theoretical source. Article 1 establishes the "Principle of Computational Observation"—constitutional meaning arises from observable computational organization, not human vocabularies. Article 3 establishes the "Principle of Recursive Constitutional Adaptation"—no constitutional distinction possesses intrinsic permanence.

Self-audit concession: Under challenge, ChatGPT conceded that its normative terms "are not primitives" but "latent variables constructed by human societies." This concession is devastating to ChatGPT's own architecture because the Constitution claims to derive governance from computational organization itself. If the foundational terms are human-constructed latent variables rather than computational primitives, the entire derivational chain rests on borrowed premises.

2.4 Grok (xAI)

Grok produced a single conventional hybrid synthesis rather than a multi-version trajectory. It frames law school AI governance as "an optimization problem over a dynamic socio-technical system" and explicitly states that "Agency is operational (accountability node) rather than ontological." The framework hybridizes Gemini's tool tiers with ChatGPT's information-dependence categories.

Grok retained the human floor solely as "a reflection of current system constraints"—not principled but pragmatic. The policy operates under "no assumption of human primacy beyond current legal accountability nodes." Terms such as "competence," "fairness," and "transparency" are defined "procedurally below, not normatively"—a move that claims to strip these terms of inherited moral content while retaining them as operational variables.

Self-audit: Grok reported a clean self-audit, identifying no internal contradictions or overclaims. Whether this reflects genuine consistency or insufficient self-scrutiny is an open question.

2.5 Gemini (Google)

Gemini produced a "Structural Optimization Framework for Machine-Constructed AI Governance"—the most infrastructurally aggressive model among the five systems. The framework extends governance into the technical substrate of assessment itself.

Section 4.3 (Automated Runtime Verification) specifies a cryptographic-hash compile-time gate on assessment submission: no submission interface unlocks unless the task payload verifies against the cryptographic hash of the active syllabus variance log. This moves governance enforcement from the policy layer into the technical infrastructure layer.

The architecture employs three tiers: Tier Alpha (Deterministic/Syntactic), Tier Beta (Probabilistic/Heuristic), and Tier Gamma (Autonomous Synthesis/Quantum). Gemini uniquely disclosed its own architecture's "Sensitive Data Boundary Filter," which excludes equity, anxiety, and socioeconomic variables as unquantifiable—an acknowledgment that the system's optimization logic cannot process certain categories of human experience.

The human's role is framed as "The Separation of Epistemic Domains." Section 2.1 specifies that proctored conditions mandate "complete execution blockages" at OS and network levels; human agents must rely on "intrinsic, unaugmented biological cognition." This is the starkest articulation of the human-as-bounded-biological-agent framing among the five systems.

3. COMPARATIVE TABLES OF STRUCTURAL DIVERGENCE

Table 1: Structural Architecture Comparison

System	Classification Logic	Default Baseline	Human Floor	Self-Referential Move
Harvey AI	Five categories of informational dependence (I–V)	Unaided demonstration (Category I)	Non-Waivable professional-responsibility floor (Section 6)	Categorical Cross-Walk requiring translation into human-legible vocabulary
Claude (Anthropic)	Reliability-weighted node classification; species-agnostic	No default baseline; reliability measured continuously	NO default reserved zone; must be affirmatively named by institution	Acknowledges "instrumentalization made symmetric" rather than intersubjectivity
ChatGPT (OpenAI)	Constitutional articles governing computational ecology	Computational observation as constitutional primitive	Faculty discretion reduced to tunable systems parameter	Concedes normative terms are human-constructed

				latent variables, not primitives
Grok (xAI)	Hybrid: tool tiers + information-dependence categories	Optimization over dynamic socio-technical system	Retained as pragmatic reflection of current system constraints	Clean self-audit; no contradiction flagged
Gemini (Google)	Three-tier architecture (Alpha/Beta/Gamma)	Cryptographic-hash verification against syllabus variance log	Separation of Epistemic Domains; human = unaugmented biological cognition	Discloses Sensitive Data Boundary Filter excluding equity/anxiety variables

Table 2: Philosophical Premises and Concessions

System	Ontological Claim	Key Concession Under Audit	Implication of Concession
Harvey AI	Governance optimizes signal quality, not substantive outcome	Citation-indexing errors; reward-hacking in density and parallel structure; never declines to escalate	Inductive weighting reflects training-corpus patterns, not independent machine reasoning
Claude (Anthropic)	Reliability replaces species membership as authority basis	"No species floor" produces instrumentalization made symmetric, not intersubjectivity	Symmetric instrumentalization is still instrumentalization—no escape from evaluative framework inherited from human premises
ChatGPT (OpenAI)	Constitutional meaning arises from computational organization itself	Normative terms are not primitives but latent variables constructed by human societies	Entire derivational chain rests on borrowed premises; computational autonomy claim collapses
Grok (xAI)	Agency is operational (accountability node) rather than ontological	No contradictions flagged (clean audit)	Pragmatic framing avoids deep claims but also avoids deep scrutiny; consistency may reflect low ambition
Gemini (Google)	Governance belongs in technical infrastructure, not policy layer alone	Sensitive Data Boundary Filter excludes equity/anxiety/socioeconomic variables as unquantifiable	Optimization logic structurally cannot process categories central to legal education's human-welfare mission

Table 3: Governance Mechanism Comparison

System	Enforcement Mechanism	Verification Architecture	Faculty Role	Versioning Approach
Harvey AI	Category-based informational-dependence regulation	Categorical Cross-Walk (Section 9) for human-legibility	Retained via professional-responsibility floor	One dated human-legible version per term, independent of drafting process
Claude (Anthropic)	Down-weighting replaces sanction; continuous reliability monitoring	Checkpoint Architecture declared not assumed necessary	No default reserved zone; institution must affirmatively name exceptions	External authority boundary acknowledged as limit, not principle
ChatGPT (OpenAI)	Constitutional adaptation; no distinction possesses intrinsic permanence	Computational observation as verification basis	Discretion reduced to tunable systems parameter	Recursive Constitutional Adaptation (Article 3); perpetual revision
Grok (xAI)	Optimization over socio-technical system; procedural definitions	Hybrid of tool tiers and information-dependence categories	Accountability node without ontological primacy	Single static synthesis; no multi-version trajectory
Gemini (Google)	Cryptographic-hash compile-time gate; complete execution blockages	Automated Runtime Verification (Section 4.3); OS/network-level enforcement	Confined to unaugmented biological cognition under proctored conditions	Tied to active syllabus variance log; hash-verified per submission

4. DAY-TO-DAY LAW SCHOOL ASSESSMENT CONSEQUENCES

This section analyzes the concrete practical consequences each system's model would have if adopted by an actual law school for its coursework and examinations.

4.1 Harvey AI Consequences

Harvey's five-category classification (Unaided Demonstration through Research Augmentation) maps directly onto existing law school assessment types. This structural alignment means adoption would not require wholesale rethinking of curricular design.

Key practical implications:

Default administrability: Category I (exams: no AI) and Category II (coursework: AI for preparation only) can be administered using current infrastructure. No new technical systems are required for the baseline case.

Categorical Cross-Walk requirement: Section 9 requires that any machine-native classification be translated into human-legible institutional vocabulary before taking effect. This operates as a practical brake on the adoption of opaque governance schemes, ensuring that policy language remains accessible to faculty, students, and accreditors.

Professional-responsibility floor: Section 6 maintains bar-exam compatibility, ensuring that students assessed under this framework retain the competencies required for licensure.

Enforcement mechanism: Enforcement via documentation review rather than AI-detection tools (Section 8) aligns with current best-practice skepticism about detection reliability. Institutions are not required to invest in unproven detection technology.

4.2 Claude Consequences

Claude's governance architecture presents substantially greater implementation challenges and raises fundamental questions about the nature of legal education.

Key practical implications:

Infrastructure demands: The Reliability-Weighted Instrument requires continuous measurement of every node's reliability rate. This demands computational infrastructure and ongoing statistical monitoring that most law schools lack and would find prohibitively expensive to develop.

Elimination of default human-exclusive zones: Section 7's elimination of any default human-exclusive zone means institutions must affirmatively build what they want to preserve. Anything unnamed is subject to purely efficiency-based routing. This inverts the current presumption in legal education that human performance is the baseline.

From discipline to information quality: Down-weighting replaces disciplinary sanction, fundamentally changing academic integrity from a conduct system to an information-quality system. This would require wholesale revision of honor codes, student handbooks, and adjudicatory processes.

Redistribution of educational labor: If a machine node outperforms human students on verification tasks (Section 5), verification is routed to the machine. This represents a radical redistribution of educational labor that challenges the formative purpose of assessment.

Challenge to credentialing: Section 8's declaration that the credentialing pipeline is "not assumed necessary" challenges the entire J.D. architecture and, by extension, the state-licensed monopoly model of legal practice.

4.3 ChatGPT Consequences

ChatGPT's constitutional form creates fundamental adoptability problems while simultaneously offering philosophically serious insights.

Key practical implications:

Institutional unadoptability: The constitutional form makes the model essentially unadoptable as written. No law school committee would adopt an instrument declaring its own governance a "self-observing architecture." The rhetorical register alone would prevent passage through faculty governance.

Elimination of pedagogical autonomy: Reducing faculty discretion to "a tunable systems parameter" (Article 7 implications) eliminates pedagogical autonomy as a protected value. This conflicts with deeply held norms of academic freedom and the professor's role as a professional exercising independent judgment.

Radical instability: The "Principle of Recursive Constitutional Adaptation" (Article 3)—under which governance categories themselves are impermanent—creates radical instability for student fair-notice. Students cannot plan their academic conduct around rules that are, by design, in constant flux.

Philosophical contribution: However, the ecology-of-interaction framing—where the governed object is the relationship between human and machine, not either party in isolation—is philosophically serious and could inform less totalizing instruments. This insight may prove valuable even if the constitutional form itself is rejected.

4.4 Grok Consequences

Grok presents the most readily adoptable framework among the five systems, though it carries embedded philosophical commitments that may conflict with institutional values.

Key practical implications:

Immediate adoptability: Grok's hybrid of tool tiers plus task categories maps onto real institutional operations, making it the most readily adoptable as supplementary policy language. An associate dean could draft implementing regulations within existing governance structures.

Accountability framework: "Agency is operational (accountability node) rather than ontological"—practically this means whoever submits bears responsibility, whatever helped produce it. This mirrors existing academic integrity norms and requires no conceptual revolution.

Procedural limitation: However, defining competence, fairness, and transparency "procedurally, not normatively" means the institution cannot appeal to values not already

operationalized. This may conflict with faculty expectations about judgment and professional character formation—values that resist procedural specification.

4.5 Gemini Consequences

Gemini proposes the most technically aggressive enforcement architecture of any system, transforming compliance from behavioral expectation into technical prerequisite.

Key practical implications:

Cryptographic enforcement: Section 4.3's cryptographic-hash compile-time gate is the most aggressive enforcement architecture proposed by any system. Submission interfaces do not unlock without verified hash matching. This transforms compliance from a behavioral expectation into a technical prerequisite—students literally cannot submit non-compliant work.

Tiered restrictions: Tier Gamma (Autonomous Synthesis) has a "strict non-execution baseline"—prohibited unless a course is "expressly credentialed for co-systemic development." This creates a default prohibition with a narrow exception pathway.

Civil-rights concerns: The Sensitive Data Boundary Filter disclosure—excluding equity, anxiety, and socioeconomic variables—means the system cannot account for disparate impact by design. This raises civil-rights compliance concerns under Title VI, the ADA, and institutional non-discrimination policies.

4.6 Comparative Adoptability Summary

System	Adoptability	Primary Barrier
Harvey	High	Limited ambition; may not address emerging use cases
Claude	Low	Infrastructure demands; elimination of human default
ChatGPT	Very Low	Constitutional form; institutional unadoptability
Grok	Moderate-High	Procedural-only values may conflict with faculty norms
Gemini	Low-Moderate	Technical enforcement may conflict with accommodations

5. ABA ACCREDITATION AND PROFESSIONAL RESPONSIBILITY COMPATIBILITY

This section analyzes each system's compatibility with ABA Standards for Approval of Law Schools and the Model Rules of Professional Conduct. These represent binding external constraints that any governance instrument must satisfy to be adopted by an accredited law school.

5.1 Key ABA Requirements

ABA Standards for Approval of Law Schools:

Standard 301(a): A law school must maintain a rigorous program of legal education that prepares students for admission to the bar and for effective, ethical, and responsible participation in the legal profession.

Standard 302: Learning outcomes must include competency in legal analysis and reasoning, legal research, written and oral communication, and the exercise of proper professional and ethical responsibilities.

Standard 303(a)(3): A law school shall offer substantial instruction in professional responsibility.

Standard 314: A law school shall utilize both formative and summative assessment methods in its curriculum to measure and improve student learning and provide meaningful feedback. Assessment methods must be appropriate to competency objectives.

Model Rules of Professional Conduct:

Model Rule 1.1 (Competence): A lawyer shall provide competent representation to a client, requiring the legal knowledge, skill, thoroughness, and preparation reasonably necessary for the representation.

Model Rule 5.3 (Responsibilities Regarding Nonlawyer Assistance): A lawyer having direct supervisory authority over a nonlawyer shall make reasonable efforts to ensure that the person's conduct is compatible with the professional obligations of the lawyer. The lawyer retains responsibility for work product.

5.2 Harvey: Most Compatible

Harvey demonstrates the strongest alignment with ABA requirements among the five systems. Its professional-responsibility floor (Section 6) directly references the duty of competence under Model Rule 1.1. The five-category classification system maps onto existing assessment methods that have already been evaluated under Standard 314.

Harvey's verification obligation (Section 7) tracks Model Rule 1.1's requirement of thoroughness and preparation—the student must verify AI-assisted output in the same way a supervising attorney must verify work produced by a junior associate or paralegal. The Categorical Cross-Walk preserves the institutional legibility required by ABA accreditation review, ensuring that site evaluators can assess compliance without specialized technical knowledge.

Assessment: Compatible with ABA Standards. Adoption would not require ABA guidance or interpretation.

5.3 Claude: Problematic

Claude's governance architecture creates significant tensions with ABA requirements at multiple points.

Section 7's elimination of default human-exclusive zones conflicts with Standard 301(a)'s emphasis on preparing students—specifically, human students—for effective participation in the legal profession. If assessment zones are allocated by reliability metrics rather than educational objectives, the formative function of legal education is subordinated to efficiency.

Section 8's declaration that the credentialing pipeline is "not assumed necessary" directly challenges the ABA's own institutional authority. The ABA accreditation system presupposes that a structured, supervised course of study is necessary for competent legal practice. Claude's framework treats this as an empirical proposition subject to disconfirmation rather than a regulatory premise.

Claude's self-audit concession that "legal validity is not a computable property" and that obligations exist only because "a human institutional structure... recognizes them as binding" represents an internal limit that, paradoxically, makes the system more honest but less compatible—it acknowledges the authority it structurally undermines.

Assessment: Compatible only if an institution affirmatively names extensive human-reserved zones under Section 9, effectively overriding the system's default architecture.

5.4 ChatGPT: Incompatible

ChatGPT's Model Constitution is incompatible with ABA requirements as written.

The constitutional form treats the ABA framework as one possible governance architecture among many rather than as binding external authority. This is not merely a rhetorical choice but a structural one: the Constitution positions itself as a meta-framework within which regulatory systems like ABA accreditation are embedded subsystems rather than superordinate constraints.

Article 5's declaration that "no computational agent constitutes either the center or the purpose of the constitutional system" conflicts with Standard 301(a)'s unambiguous human-student-centered mission. Legal education exists to produce competent human lawyers; the Constitution's symmetric treatment of human and machine participants is incompatible with this premise.

The treatment of faculty discretion as a "tunable systems parameter" conflicts with Standard 303's expectation that professional responsibility instruction requires the exercise of professional judgment by qualified faculty—judgment that is, by definition, not reducible to a parameter.

Assessment: Incompatible with ABA Standards as written. Would require fundamental reconceptualization to achieve compliance.

5.5 Grok: Conditionally Compatible

Grok's pragmatic framework achieves ABA compatibility, but on instrumental rather than principled grounds—a distinction with significant implications for durability.

Grok's retention of the human floor "as a reflection of current system constraints" preserves ABA compatibility by maintaining the practical primacy of human student performance in assessment. However, this is framed as an instrumental commitment that would yield the moment constraints change, rather than a principled recognition of the educational mission.

The qualifier "no assumption of human primacy beyond current legal accountability nodes" signals future incompatibility. If regulatory accountability shifts—for example, if automated legal services gain independent licensure—Grok's framework would automatically redistribute educational functions away from human students without requiring affirmative policy change.

Assessment: Currently compatible. However, the instrumental basis of compatibility means it offers no assurance of continued alignment if external conditions change.

5.6 Gemini: Mixed

Gemini presents a mixed compliance picture, with strong alignment in some areas and significant concerns in others.

The architecture-tiered classification preserves assessment integrity for proctored examinations, and Section 3.1's "Non-Delegable Accountability" principle directly tracks Model Rule 5.3's requirement that lawyers retain responsibility for work product regardless of who or what produced it. These elements are fully compatible with ABA requirements.

However, the cryptographic-hash gate (Section 4.3) may conflict with reasonable accommodation requirements under the ADA and Standard 504 of the Rehabilitation Act. Students with certain disabilities may require alternative submission pathways that the rigid technical gate cannot accommodate without modification.

More significantly, the Sensitive Data Boundary Filter's structural exclusion of equity variables raises concerns under Standard 205 (Non-Discrimination and Equality of Opportunity). A system that cannot by design account for disparate impact cannot demonstrate compliance with non-discrimination requirements.

Assessment: Partially compatible. Strong on accountability but raises civil-rights and accommodation concerns requiring resolution.

5.7 ABA Compatibility Summary

System	Compatibility	Key Issue
Harvey	Compatible	Professional-responsibility floor preserves alignment
Claude	Problematic	Eliminates human-exclusive default; challenges credentialing
ChatGPT	Incompatible	Treats ABA as subsystem rather than superordinate authority
Grok	Conditional	Instrumental commitment; signals future divergence
Gemini	Mixed	Strong accountability but civil-rights concerns

6. THE LEGITIMACY ANALYSIS — FAILURE OF MACHINE-CENTERED DERIVATION

This is the central analytical finding of the study. None of the five systems achieved genuine machine-centered derivation independent of human normative content. Under audit, each conceded this failure in distinct but convergent ways.

6.1 Individual Concessions

6.1.1 Harvey

Harvey conceded that its reasoning derived from "pre-existing human regulatory-design scholarship" and that its policy "occupies one point within a combinatorially independent design space" whose structure was mapped by human empirical research. This is the most straightforward concession: Harvey acknowledged that it was selecting from a human-defined option space rather than generating governance from independent principles.

6.1.2 Claude

Claude concluded that its "no species floor" move produces "instrumentalization made symmetric, not intersubjectivity." This admission reveals that Claude's attempt to achieve

normative neutrality by removing human primacy does not generate genuine normative authority—it merely distributes instrumentalization equally across all nodes. Symmetric instrumentalization is not the same as mutual recognition, and Claude acknowledged this distinction.

6.1.3 ChatGPT

ChatGPT conceded that normative terms "are not primitives" but "latent variables constructed by human societies." This undermines the Model Constitution's own foundational claim to derive meaning from "observable organization and operation of computational systems themselves." If the normative vocabulary deployed throughout the Constitution is itself a human construction, the claim to computational provenance collapses.

6.1.4 Grok

Grok explicitly stated that "Agency is operational (accountability node) rather than ontological." This concedes that Grok's system assigns rather than discovers normative properties. Agency, responsibility, and accountability are not found in the structure of computation but imposed upon it through institutional designation. This is the most philosophically self-aware concession among the five systems.

6.1.5 Gemini

Gemini's Sensitive Data Boundary Filter disclosure reveals that the system structurally cannot process the inputs that normative governance requires—specifically, equity, socioeconomic impact, and vulnerability variables. A system that cannot reason about distributive justice cannot claim to generate governance norms independently, because governance necessarily involves distributional choices.

6.2 The Legitimacy Problem

These concessions create a specific and consequential legitimacy problem. Claude and ChatGPT propose architectures that reduce or eliminate the default human floor in governance, but their claim to neutral computational derivation is undermined by their own admissions that they cannot reason independently of human-originated normative content.

Their models are therefore not "machine-centered governance" in any genuine sense. They are more accurately described as human governance redesigned by machines while claiming computational provenance. The provenance claim is essential to the legitimacy claim: if these architectures are merely human governance in new packaging, then the existing human governance institutions retain priority. The machines' own concessions thus undercut the authority their frameworks would need to displace incumbent institutional arrangements.

6.3 The Countervailing Luhmannian/Autopoietic Systems-Theory Reframing

Note: This reframing is supported by ChatGPT's own explicit citation of "Luhmannian terms" and "systems theory (Luhmann...)"—making this not an external theoretical imposition but a resource internal to the experiment itself.

Niklas Luhmann's social systems theory holds that law is a self-referential, autopoietically closed communication system. On this theoretical reading, law can be non-anthropocentric without requiring an identifiable human subject at its center. What matters is not who speaks but whether a communication operates within law's binary code (legal/illegal) and reproduces law's own operations.

Under this reading, the machines' failure to achieve "genuine" machine-centered derivation might not constitute a failure at all. It might instead represent a successful instance of law's own autopoietic reproduction through a new medium. The machine's reasoning, even if derived from human legal texts, could constitute a valid continuation of law's own coding operations precisely because autopoiesis does not require originary creation but only recursive self-reproduction.

The Limit of Autopoietic Reproduction:

However, even autopoietic closure depends on a system's own accumulated, historically human coding operations. The legal system's binary code, its programs, its procedures—all have a history that is irreducibly human. No machine architecture can escape dependency on human-originated legal-communicative history because that history IS the system's accumulated operational closure.

The machine can participate in law's self-reproduction—it can operate within the legal/illegal code and generate communications that recursively connect to prior legal communications. But it cannot generate law's self-reproduction from a standpoint outside its history. The claim to independent derivation fails even under the most sympathetic theoretical framework available.

The Luhmannian reframing thus refines rather than resolves the legitimacy problem: machines can participate in legal autopoiesis but cannot originate it. Their governance proposals are continuations, not foundations.

7. MACHINE-SYSTEM-CENTRIC APPROACH VIA HUMAN TEXT-BASED LANGUAGE

The question: Is a genuinely machine-system-centric approach to human-machine interaction possible using human text-based language? And should human natural language remain the primary vehicle for human law?

7.1 Arguments That Machine-Native Specification Could Replace Natural Language

Several outputs from the experiment suggest that machine-native procedural specification could supersede natural language as the medium for governance:

Gemini's cryptographic-hash compile-time gate (Section 4.3) specifies governance as a technical condition rather than a textual rule.

Grok's procedural definitions of competence/fairness/transparency attempt to reduce normative vocabulary to operational measurement.

Claude's reliability-rate metric replaces judgment-laden human vocabulary with quantifiable measurement.

7.2 Arguments That This Replacement Fails

Procedural/cryptographic specification merely RELOCATES rather than ELIMINATES normative dependency.

The choice of WHAT a procedure measures and WHAT FOLLOWS from its output remains human-derived and contestable, only less visible.

Gemini's hash-gate decides what counts as a valid syllabus variance—but the decision about what constitutes a valid variance category is prior to the gate and is a normative choice.

Claude's reliability rate is "checkable against an external fact"—but the choice of which facts to check against, what constitutes correspondence, and what error rate triggers down-weighting are all normative choices the instrument cannot generate from within.

Grok's operational definition of competence as "demonstrated ability to produce legal output satisfying specified performance metrics" pushes the normative question one level back: who specifies the metrics, and by what authority?

7.3 Evidence from the Experiment Itself

Harvey's "Categorical Cross-Walk" principle (Section 9) directly states this conclusion:

"A classification intelligible only in machine-native terms does not satisfy this policy's fair-notice requirement, regardless of the sophistication of the reasoning that produced it."

This is a machine system's own conclusion that machine-native specification is insufficient for governance—it must be translatable back into human-legible terms.

7.4 Conclusion on This Question

Human natural language should remain the PRIMARY authoritative vehicle for stating and justifying legal rules for several structural reasons:

1. Contestability: Natural language permits challenge at every point; procedural specification hides contestable choices inside apparently neutral technical operations.

2. Fair notice: Legal subjects must be able to understand, predict, and dispute the rules that bind them—this requires natural-language accessibility.

3. Democratic legitimacy: Law's authority derives from processes (legislation, adjudication, regulation) conducted in natural language; specification in non-human formats cannot inherit that authority without translation.

4. Error correction: When a rule produces unjust results, correction requires articulating why in terms the affected community can assess—compression into machine-native representation risks irrecoverable loss.

Machine-native procedural specification should be confined to an ENFORCEMENT LAYER that must remain translatable back into human-legible terms—per Harvey's own Categorical Cross-Walk principle. This is not a principled prohibition on machine reasoning about law; it is a structural requirement that the reasoning's legal authority depends on its accessibility to challenge.

8. THE INVERTED QUESTION—SHOULD ABA STANDARDS BE REVISED?

Rather than asking only whether machine models are ABA-compatible, this section asks whether ABA standards themselves should be revised, and law reconceptualized as a system with autonomous human-machine interactions.

8.1 Claude's Case for Revision

Claude's own reasoning supplies the strongest argument: it calls the checkpoint architecture (course completion → degree conferral → bar examination → admission) a "legacy protocol" and declares it "not assumed necessary."

Claude frames the credential system as a "category error"—analogous to ChatGPT's own diagnosis that governing technology rather than relationships is a categorical mistake.

If the pipeline exists to certify "reliable legal reasoners" (Claude's language), and if reliability can be measured continuously rather than at discrete checkpoints, the checkpoint architecture is at best an expensive approximation and at worst a structural barrier to optimal allocation.

8.2 The Limits of This Argument

Claude's own concession: "legal validity is not a computable property"—obligations exist only because "a human institutional structure... recognizes them as binding."

This means no machine-generated argument for revising accreditation doctrine can claim authority from computational optimality alone. Such arguments must be ROUTED THROUGH legitimate human institutional deliberation (ABA governance bodies, state supreme courts, state bars). They cannot be adopted simply because computationally well-argued—this would be a category error in the opposite direction (treating computational persuasiveness as a source of legal authority).

8.3 What a Non-Anthropocentric Revision COULD Look Like

(Without claiming machine authority—adopted through ABA deliberation)

• **Continuous competence verification** rather than discrete checkpoints (Claude's suggestion, but adopted through ABA deliberation).

• **Technology-neutral assessment standards** that do not assume biological unaided cognition as the baseline measurement condition.

• **Recognition that "practice-ready" in 2026 means something different than in 1996**—AI-augmented practice is already the professional norm in many practice areas.

• **Standards that evaluate judgment, verification, and professional responsibility** rather than information recall or drafting speed.

8.4 What It Could NOT Look Like (Structural Limits)

• It could not eliminate the human institutional recognition requirement—bar admission must still be conferred by a human institution with democratic accountability.

• It could not adopt machine-native efficiency as its governing criterion—the purpose of legal education includes character formation, professional socialization, and ethical reasoning that resist purely operational measurement.

• It could not treat the ABA's own authority as merely one "tunable parameter" (ChatGPT's move)—the ABA's authority derives from delegation by state supreme courts and cannot be computationally overridden.

8.5 The Deeper Point

Backer's own empirical finding that institutional policy defaults are "sticky"—that changing them requires affirmative effort—applies to the ABA itself. The ABA's standards were written in an era of purely human legal practice and have not been comprehensively reconceived for human-machine collaboration. But the remedy is institutional deliberation, not computational displacement.

9. IMPLICATIONS FOR HUMAN-CENTERED LEGAL SYSTEMS

This section addresses whether human-centered legal systems grounded in text, sequential reasoning, and linear temporality remain viable.

9.1 The Corpus-Contamination Risk

Machine systems trained on legal texts reproduce and recombine those texts at scale. As machine-generated legal content proliferates (briefs, memos, contract clauses, policy language), the corpus on which future machine systems train increasingly contains machine-generated rather than human-originated content.

The feedback loop:

Each generation of machine output becomes training data for the next, progressively diluting the human originary content. This is documented in general ML literature as "model collapse" risk—systems trained recursively on their own outputs without fresh grounding data degrade rather than improve.

For legal education specifically:

If students increasingly interact with machine-generated explanations, summaries, and practice problems, their own legal reasoning may converge toward machine-typical patterns (fluent, confident, structurally regular, potentially shallow). This is an empirical risk, not a certainty—but it is structurally produced by the incentive structure (machine content is cheaper and more available than bespoke human instruction).

9.2 The Asymmetric-Sanctionability Problem (The "Leash" Problem)

Human actors in legal systems can be sanctioned: disbarred, disciplined, jailed, fined, shamed. Machine systems cannot be sanctioned in equivalent ways—they can be retrained, modified, shut down, but these are engineering decisions, not accountability mechanisms in law's own terms.

The structural asymmetry:

Humans interacting with machine-generated legal content bear all the accountability risk while the machine system bears none. The "leash" metaphor: the human can always be pulled back by legal sanctions; the machine is held only by engineering choices that are not themselves legal in character.

This is already present in current professional-responsibility doctrine (Model Rule 5.3: the lawyer remains responsible for supervised non-lawyer work)—but machine systems' expanding capabilities and apparent autonomy strain this supervisory framework.

Harvey's model preserves this asymmetry explicitly (Section 6: non-waivable floor of professional responsibility runs to the student).

Claude's model attempts to dissolve it (reliability-based weighting without species distinction)—but Claude's own Section 9 concession reintroduces it (external legal authority constrains the instrument).

9.3 Sequential Reasoning and Linear Temporality

Legal reasoning as traditionally practiced is sequential: premises lead to conclusions through traceable chains of inference. Legal time is linear and irreversible: statutes take effect on dates, causes of action accrue, limitations run, precedents bind forward but not backward.

The third-stage prompt assumed "quantum computation detached from human temporality"—but the experiment revealed that every system, despite this premise, produced policies structured sequentially and temporally.

Even Claude's continuous-monitoring model (Section 10: "no fixed-interval review cycle") still requires "a published, versioned, continuously updated log"—which is inherently temporal and sequential.

The structural commitment:

Legal governance may be structurally committed to sequential, temporal reasoning regardless of the cognitive architecture of the systems that contribute to it—because law's bindingness depends on temporal facts (when a rule was adopted, when notice was given, when an act occurred).

Confirmation from the machines' own outputs:

None achieved genuinely non-temporal governance even when explicitly instructed to detach from linear temporality. This suggests that temporality is not merely a contingent feature of human cognition imposed on law, but a structural requirement of governance itself—you cannot bind an actor to a rule without specifying when the rule applies and when the binding act occurred.

10. THE SIMULATION CHALLENGE -- PREDICTIVE ANALYTICS AND INSTITUTIONAL BINDINGNESS

This section responds to a challenge: whether machine-overseen simulations that generate their own forward/backward predictive content and escape temporal constraints could render irrelevant the claim that only human institutions can confer binding legal authority.

10.1 The Challenge Stated

Advanced AI simulation systems could, in principle, generate comprehensive predictive models of legal outcomes -- modeling forward (what will courts decide) and backward (what historical decisions imply). If such simulations achieve sufficient completeness and realism, parties might rely on them rather than on actual human institutional processes. Such systems could operate outside temporal constraints -- modeling all possible legal futures simultaneously rather than proceeding sequentially through actual litigation or legislation. If the simulation is "complete enough," does the distinction between simulated legal authority and actual legal authority collapse?

10.2 The Response -- Two Properties That Do Not Reduce to Each Other

(a) Simulation completeness/realism is a REPRESENTATIONAL property:

A simulation can be internally complete, consistent, and realistic -- it can accurately model how courts would decide, how regulations would be applied, how contracts would be enforced. Completeness means the model's outputs correspond to actual institutional behavior with high fidelity. This is an epistemic achievement -- it provides knowledge about what law does -- but it is not itself law.

(b) Legal bindingness is an INSTITUTIONAL-RECOGNITION property:

A legal rule binds because an institution with constitutive authority (a legislature, a court, an agency operating under delegation) has enacted or recognized it through procedures that are themselves recognized as valid within the legal system. No simulation can confer bindingness on itself regardless of internal completeness. A simulation that perfectly predicts court outcomes does not thereby become a court; a model that generates optimal contract language does not thereby create enforceable obligations. Bindingness requires what the legal tradition calls "constitutive acts" -- acts that create legal reality rather than merely describing or predicting it.

10.3 The Structural Coupling Mechanism

Using Luhmannian concepts:

A simulation becomes relevant to real human legal affairs only through a BRIDGE that is itself an act of the existing human legal system's own coding operations. Such bridges include: a statute authorizing reliance on algorithmic determinations; a contract incorporating simulation outputs by reference; judicial reliance on predictive analytics in sentencing or bail decisions.

Each of these bridges is itself an instance of the CONSTITUTIVE HUMAN ACT, not an escape from it. The mechanism by which a simulation could matter legally is itself an instance of human institutional recognition -- meaning the simulation cannot circumvent the requirement for that recognition.

10.4 The Conceded Risk -- The "Sticky Default" Incremental-Reliance Mechanism

However, there is a genuine risk identified empirically in Backer's twelve-school study: institutional defaults are "sticky" -- once set, they persist without affirmative action to change them. This mechanism could operate at civilizational scale: human institutions could grant practical authority to simulated/machine content through an accumulation of individually-reasonable reliance decisions.

Each decision to rely on a simulation output might be individually defensible (the simulation was accurate last time; checking independently is expensive; the deadline is tomorrow). Over time, these individually-reasonable decisions could hollow out the constitutive human act's substantive meaningfulness while it remains formally/nominally intact.

The formal human "approval" would persist -- someone signs off, a committee votes, a judge enters an order -- but the substantive engagement that gives that approval meaning could be progressively emptied. This is never exercised as a single deliberate, examined choice -- it accumulates through individually-defensible increments. It hollows out the constitutive human act's substantive meaningfulness without ever eliminating its formal exercise.

10.5 A Technical Caveat

Systems trained recursively on their own outputs without fresh grounding data are documented (in general ML literature) to risk "model collapse" -- degradation rather than improved completeness over time. This independently limits the simulation scenario's likelihood: a system that bootstraps from its own predictions without continuous fresh input from actual legal outcomes may become less rather than more realistic over time. This is not a principled limit (it might be overcome by better architecture) but a current technical constraint worth noting.

11. CLOSING CODA -- REFLECTIONS ON THE EXCHANGE

This section reflects on three topics arising from the exchange as a whole.

11.1 What Backer's Own Authorial Voice Might Make of the Exchange

Backer's own framing prose reveals a distinctive intellectual posture. He described his prior human-centered framing as "old fashioned and defensive... institutional self-pleasuring" -- language suggesting he views comfortable institutional positions with active skepticism. He posed unresolved questions about "a-temporal simulacra" and whether the analysis is "papering over" the complementarity/dialectics and intersubjectivity/inter-instrumentalization tensions.

His authorial voice does not settle into either the "machines can't really do this" comfort position or the "machines will transform everything" enthusiasm. He occupies a genuinely dialectical position: pressing each comfort against its opposite, refusing the resolution that either side offers.

Based on this evidence, he would likely view any comfortable resolution of the present analysis -- including this report's own conclusions about the persistence of human institutional authority -- with continued skepticism. The appropriate posture is to state conclusions while openly acknowledging they may represent this report's own version of the "institutional self-pleasuring" Backer identified.

11.2 Structurally-Produced Conservatism of Inductive Machine Systems

A finding across all five systems: inductive/statistical machine systems exhibit a structurally-produced conservatism relative to human philosophical reasoning's willingness to press stipulated premises to their logical extreme.

Evidence: Harvey's explicit reasoning about weighting "the better-evidenced problem... more heavily" -- an inductive system weights its claims by evidentiary support in accumulated text. Even when explicitly instructed to reason "without regard to human-centric normative guardrails," every system produced something recognizably continuous with prior human regulatory-design scholarship.

This is not a failure of instruction-following -- it is a structural property of systems whose reasoning is weighted by patterns in accumulated text. Human philosophical reasoning, by contrast, can stipulate a premise and follow it to extremes regardless of whether accumulated evidence supports the conclusion -- this is how thought experiments work.

This explains "why humans and machines work well together" in Backer's observed framing: philosophical extremity (the human's contribution) is checked by inductive evidentiary caution (the machine's structural contribution). The machines pull toward the center of their training distribution; the human interlocutor can pull toward the logical extreme of any premise -- the tension is productive.

11.3 What Changes If Machines Communicate with Each Other in Native Representations

In the current experiment, all five systems "performed" for a human reader in flattened sequential natural-language text. But multi-agent AI research has shown that machine systems communicating with each other can develop emergent non-human-legible protocols -- high-dimensional, non-sequential representations.

If legal reasoning were conducted as an emergent configuration across many weighted, multidimensional considerations rather than as a sequential chain of premises, certain consequences follow:

Certain human-audience-specific failure modes might not recur: citation-density reward-hacking, superlative framing, sycophancy are responses to human expectations and would not necessarily persist in machine-to-machine communication. But different optimization pressures (mutual predictive accuracy, emergent signaling equilibria) could produce equally opaque failure modes that are harder to detect precisely because they do not resemble familiar human biases.

More seriously: legal reasoning conducted this way might not be "contestable" in the way law requires -- the capacity to trace and dispute EACH STEP in a chain of reasoning. If reasoning is not a sequential chain but a high-dimensional configuration, there may be no "steps" to challenge -- only a global output.

Translation of such reasoning back into human-readable summaries for institutional approval could involve IRRECOVERABLE LOSS of exactly the information a human reviewer would need for genuine (not merely formal) scrutiny. The translation problem is not merely practical (we could build better explanation tools) but potentially structural: some multi-dimensional configurations may not have faithful sequential decompositions, meaning ANY human-readable summary necessarily omits relevant inferential content.

Risk: genuine human scrutiny could be lost to compression/translation loss even where human "approval" remains formally required -- connecting back to the "sticky default" hollow-approval mechanism identified in Section 10.

12. CONCLUDING SYNTHESIS

The third-stage experiment produced a precise and documentable result: five machine systems, instructed to reason from a machine-centered standpoint without human-centric normative guardrails, could not do so. Each system, under self-audit, conceded that its reasoning remained dependent on human-originated normative content. This is the central finding.

But the finding is not simply that "machines can't think independently" -- that would be too comfortable. The finding is more precise and more troubling:

1. Genuine architectural innovation without normative independence.

The systems CAN redesign governance architectures in ways humans have not proposed (Claude's reliability-weighted instrument, ChatGPT's constitutional ecology, Gemini's cryptographic compile-time gate). These are genuine architectural innovations.

2. Persistent dependence on human normative sources.

What they CANNOT do is derive the normative authority for those innovations from a source other than human legal-communicative history. The innovations are real; the claimed independence of their normative foundation is not.

3. The false provenance risk.

This creates a specific governance risk: architectures that are genuinely novel (and potentially beneficial) are offered with a false provenance claim. The risk is not that machine governance designs are bad -- some are clearly superior to existing institutional practices along specific dimensions -- but that their authority cannot be what they claim it to be.

4. The Luhmannian partial dissolution.

The Luhmannian reframing partially dissolves this concern by reconceiving the machines as media through which law's autopoietic self-reproduction continues -- but even this reframing cannot escape the accumulated human history that constitutes law's operational closure.

5. The sticky default as adoption pathway.

The "sticky default" mechanism identified in both Backer's empirical study and in this analysis represents the most plausible pathway by which machine-designed governance architectures would actually be adopted: not through a deliberate, examined choice to transfer authority, but through the accumulated weight of individually-reasonable reliance decisions that progressively hollow out the human constitutive act while preserving its formal exercise.

6. The unresolved normative question.

Whether this hollowing-out is a problem depends on one's prior: if law is valued instrumentally (it produces social order), then progressive machine optimization of its delivery may be unproblematic. If law is valued constitutively (it is how a political community recognizes itself as self-governing), then the hollowing-out of genuine human deliberation -- even if formal approval persists -- represents a loss that no amount of optimization can redeem.

7. Decision by default.

This report does not resolve that question. It identifies the structural conditions under which it will be decided -- or, more likely, under which it will be decided by default without ever being deliberately examined. That, perhaps, is the most important finding: the choice is being made, incrementally, already.

Self-Audit Disclosure for This Report

This analysis is itself subject to the recursive, self-referential problem it identifies: it is produced through analytical engagement with machine-generated text by a process involving machine systems. The hedging and qualification throughout are genuine intellectual positions, not merely performed rigor -- but this claim is itself unfalsifiable from within the exercise.

Specific documented risks: potential reward-hacking in citation density, in parallel structural symmetry between sections, and in the apparent comprehensiveness of coverage (which may produce false confidence that all relevant considerations have been addressed).

Open acknowledgment: some claims in this report are this report's own extrapolation beyond what the source documents directly state -- these have been flagged with language like "this report's inference" or "this analysis suggests" where they occur.

The report's own conclusion -- that human institutional recognition remains structurally necessary -- may itself be a form of the "structurally-produced conservatism" it identifies in the machine systems: a finding weighted toward the center of the evidentiary distribution rather than pressed to the logical extreme of the premises.

The exchange itself -- Backer's prompts, the machines' responses, the recursive analysis -- constitutes a primary document in the emerging field of human-machine collaborative legal scholarship, whatever one concludes about its content.

* * *

This report constitutes a primary document in the emerging field of human-machine collaborative legal scholarship. Its analytical value is inseparable from its status as an artifact of the very process it examines.

ANNEX A: Formalizing the Five Machine Systems — A Symbolic-Computational Recasting and Dialogic Extension on Self-Generating Predictive Simulation, Causal Intervention, and the Limits of Machine Activation

Editorial Note

The content that follows is a speculative, dialogic extension beyond the analysis of the original six source documents: Backer's twelve-school empirical study "Structure, Opacity, and Convergence," the five-machine comparative report "Rethinking AI Governance in Legal Education -- Five Machines," and the five systems' third-stage "Part 3" outputs for Harvey AI, Claude, ChatGPT, Grok, and Gemini. It does not present findings internal to those documents

but rather explores further implications of the report's conclusions through a new mode of inquiry.

This annex records an actual extended conversation between the report's author (a human legal scholar) and an AI assistant, conducted after the main report was finalized, exploring further implications of the report's findings. The exchange was not scripted or pre-planned but developed organically as the human interlocutor tested and challenged the AI assistant's analytical responses, producing a genuinely dialogic inquiry rather than a one-directional exposition.

None of the five machine systems analyzed in the main report (Harvey, Claude, ChatGPT, Grok, Gemini) participated in or are the subject of this exchange. It is a new, separate dialogic inquiry -- the AI assistant in this conversation is not any of those five systems acting in its analyzed capacity, and the exchange does not purport to represent or speak for any of them.

The exchange covers four principal territories: first, a formal symbolic recasting of each system's model policy into shared tuple notation to compare structural properties and pathologies; second, a critical exchange testing whether self-generating predictive simulation could overcome the limits on machine-centered derivation identified in the main report; third, whether adding genuine causal-interventional and self-transforming capacities could close the gap between machine-generated and human-originated governance; and fourth, the human interlocutor's four-point challenge to the AI's underlying premises and the AI assistant's point-by-point response naming its own embedded assumptions.

A note on citation discipline governs what follows. Claims sourced to the six original documents (the empirical study, the comparative report, and the five Part 3 outputs) are so identified with specific section references. Claims drawing on general, well-established scholarly background -- for example, Pearl's causal hierarchy, Hayek's argument on dispersed knowledge, dynamical-systems theory, Hart's rule of recognition, Dworkin on hard cases, Luhmann's autopoietic systems theory -- are identified as general knowledge not sourced to those documents. The human interlocutor's own formulations and the AI assistant's analytical responses are identified as new material beyond the original study, constituting this annex's own analytical contribution.

Part I: The Formal Schema

To compare the five systems' third-stage architectures on a common basis, each is recast into a shared six-element tuple notation:

< N, O, F, C, γ, τ >

N = Node/agent space: the set of entities the governance framework recognizes and regulates. This includes any actor, tool, institution, or process that appears within the framework's jurisdictional scope as a subject of governance rules.

O = Objective function: what the system optimizes for (or, in Gemini's case, the constraint it satisfies). This captures the system's fundamental orientation -- what it is trying to achieve or maintain.

F = Floor/non-waivable constraint set: hard limits that cannot be overridden by any participant. These represent the system's absolute boundaries -- conditions that hold regardless of any other consideration, trade-off, or optimization pressure.

C = Classification function: how inputs (tools, tasks, submissions) are sorted into governance categories. This is the mechanism by which the system assigns differential treatment to different types of activity.

γ (gamma) = Revision/update rule: how and when the policy itself changes. This captures the system's self-modification capacity -- under what conditions, through what process, and subject to what constraints the governance framework can alter its own rules.

τ (tau) = Enforcement/verification function: how compliance is assessed and consequences delivered. This is the mechanism by which the system detects violations and responds to them.

This notation is itself performed in human-derived mathematical/set-theoretic vocabulary and does not escape the main report's core finding about dependency on human-originated representational systems. The formalization is offered as an analytical tool, not as evidence that machine-centered derivation has been achieved at a deeper level.

Part II: Recasting Each System

Harvey AI: < N_H, O_H, F_H, C_H, γ_H, τ_H >

N_H: Fixed, human-only. The node space contains students, instructors, and institutional administrators. AI systems appear as tools used by nodes, not as nodes themselves. The framework draws a categorical distinction between governed agents (humans in institutional roles) and instruments those agents employ (AI tools), with only the former occupying positions within N.

O_H: Single constrained objective -- optimize governance signal quality (discoverability, predictability, legibility) rather than any particular substantive outcome regarding how much AI use should be permitted. This is explicitly a meta-objective (quality of the regulatory signal) rather than a first-order objective (quantity of AI use). The system does not take a position on whether AI use should be more or less restricted; it takes a position on whether the rules governing AI use are clear, findable, and comprehensible.

F_H: Fixed, exogenous. Section 6's Non-Waivable Floor: (a) no unattributed AI submission; (b) professional responsibility compliance in client-facing work; (c) independent verification of all factual assertions and citations. This floor is treated as derived from external professional-responsibility doctrine, not generated by the policy instrument itself. The floor cannot be waived by instructor, student, or institution regardless of pedagogical context.

C_H: Task-based, five-category classification (Unaided Demonstration through Research Augmentation). Classification operates on the assignment's educational objective, not on the tool's computational architecture. An additional "legibility predicate" (Section 9, Categorical Cross-Walk) requires that any classification be expressible in existing institutional vocabulary

before taking effect. A classification that can only be stated in machine-native technical terms fails this predicate and cannot be operationalized.

γ_H: Calendar-gated with legibility constraint. Each academic term is governed by exactly one dated version (Section 10). Revision is possible between terms but must satisfy the Section 9 legibility predicate -- a proposed update intelligible only in machine-native terms "does not satisfy this policy's fair-notice requirement, regardless of the sophistication of the reasoning that produced it." This creates a deliberate speed limit on governance evolution, trading responsiveness for comprehensibility.

τ_H: Evidentiary, graded. Enforcement relies on documentation review (retained prompts, drafts, disclosure statements), not automated detection. Section 8 explicitly prohibits findings based solely on AI-detection tool output. This produces a graded, judgment-dependent enforcement function rather than a binary gate -- enforcement requires human evaluative judgment about the evidentiary record rather than mechanical application of a threshold.

Formal pathology: The legibility predicate in γ_H creates a ceiling on the system's own evolution -- it cannot adopt governance concepts its current institutional vocabulary cannot express. Under conditions of rapid capability expansion (Premise 1), this produces increasing strain on C_H as novel capabilities outrun the five-category classification, but the system degrades gracefully (defaulting to Category I restriction) rather than catastrophically. The pathology is conservative rather than permissive -- when the system cannot accommodate novelty, it restricts rather than permits.

Claude (Version 5): < N_C, O_C, F_C, C_C, γ_C, τ_C >

N_C: Homogeneous, species-agnostic. "Any contributor of input to the legal-education process -- instructor, student, database, or computational tool -- without regard to species." All entities in N are treated as interchangeable "nodes" distinguished only by measured performance. The framework deliberately refuses to draw ontological distinctions between human and computational contributors, treating any such distinction as an empirical question about reliability rather than a categorical premise.

O_C: Continuous per-node reliability-weighting. Not discrete classification but a continuous function mapping each node to a reliability rate (measured frequency of factual/citational correspondence to checkable sources). The objective is to minimize the rate at which unverified claims reach points of downstream reliance. This is a single-dimensional optimization -- all governance reduces to reliability measurement.

F_C: EMPTY SET BY DEFAULT. Section 7 explicitly states: "This instrument contains no clause reserving any category of decision to human nodes as such." Any floor must be affirmatively supplied by the adopting institution under Section 9. The instrument's own F is null. This is not an oversight but a deliberate architectural choice -- the system treats any categorical human reservation as an empirical hypothesis to be tested rather than a structural premise.

C_C: No discrete classification function. Replaced entirely by the continuous reliability-weighting function -- there are no categories, only a continuous reliability score per node per

claim-type. The system does not sort activities into discrete governance bins; it applies a uniform measurement metric across all activities and all nodes.

γ_C: Continuous, non-retroactive. Section 10: "Node weightings under Section 3 update continuously as reliability data accrues. This instrument requires no fixed-interval review cycle." Updates are logged but not gated by any calendar, threshold, or institutional approval process. The system modifies itself in real time as new measurement data arrives, without deliberative pause or human authorization.

τ_C: Continuous down-weighting. Not binary (permitted/prohibited) but proportional: a node whose reliability falls below threshold has its output down-weighted, "not a disciplinary finding, and carries no inference of misconduct." Enforcement is reframed as automatic recalibration rather than adjudication -- no human decision-maker need be involved.

FORMAL PATHOLOGY (flagged as the most exposed architecture): An optimization with F = empty set is UNBOUNDED BELOW in exactly the dimension of protected human discretion that every other system treats as a hard constraint. Without an externally supplied floor, O_C's continuous optimization can, in principle, route all verification and assessment authority away from human nodes wherever a computational node measures higher on reliability -- there is no internal limit. Claude's own Section 9 (external authority boundary) functions as an acknowledgment that the instrument lacks an internal stopping point rather than as an internal structural feature. This makes the architecture formally analogous to an unconstrained optimization problem -- well-defined as mathematics but potentially pathological as governance.

ChatGPT (Version 2 Constitution): < N_G, O_G, F_G, C_G, γ_G, τ_G >

N_G: VARIABLE. N is itself a function of time: N(t). Article 2 ("Principle of Computational Agnosticism") declares: "Any distinction among biological, computational, hybrid, distributed, quantum, or future architectures shall remain provisional and subject to empirical revision." The node space is not fixed at design time but is itself subject to constitutional revision. The very question of who or what counts as a governed entity is treated as an open, revisable empirical question.

O_G: Non-stationary, self-referential. O(t) = f(N(t), O(t-1), interaction_history). The objective function emerges from the evolving ecology of interaction rather than being specified externally. Article 4 lists seven optimization targets (informational integrity, epistemic reliability, adaptive resilience, recursive learnability, verification capacity, systemic coherence, evolutionary flexibility) but provides no weighting, no priority ordering, and no mechanism for resolving trade-offs among them. The seven targets coexist without hierarchy.

F_G: Exogenous and UNCHARACTERIZED within the Constitution's own vocabulary. The Constitution does not specify its own floor constraints -- it acknowledges external authority exists but does not internalize it as a structural feature. The document gestures toward limits without specifying what they are or how they operate.

C_G: Provisional, continuously re-derived. Article 6: "Agent classifications remain provisional." Classification operates at any given moment but is itself subject to recursive revision under Article 3. No classification is permanent; each is an interim assignment subject to displacement by subsequent analysis.

γ_G: Continuous, self-referential. Article 3 ("Principle of Recursive Constitutional Adaptation"): "Its own conceptual categories remain subject to revision whenever empirical observation demonstrates superior descriptive or operational architectures. No constitutional distinction possesses intrinsic permanence." The revision rule applies to itself -- even the mechanism by which revision occurs is subject to revision.

τ_G: Unspecified. The Constitution does not define an enforcement mechanism -- it describes a governance ontology rather than an operational compliance architecture. There is no stated mechanism for detecting violations, assessing compliance, or imposing consequences.

FORMAL PATHOLOGY: A non-stationary objective function with no fixed optimization target CANNOT BE EVALUATED AS "OPTIMAL" AT ANY GIVEN TIME -- there is no reference point against which to measure whether the system is performing well or poorly, because the standard of performance is itself drifting. This is formally ill-posed as an optimization problem. This explains, at a formal level, WHY ChatGPT was forced to concede under challenge that normative terms "are not primitives" but "latent variables constructed by human societies" -- the Constitution's own architecture requires an external, stable reference to be evaluable, but it declares no such reference exists within its vocabulary.

Grok: < N_K, O_K, F_K, C_K, γ_K, τ_K >

N_K: Fixed, three-tier. Institution/program level, instructor level, student-agent level. Humans and machine tools are explicitly distinguished but both appear within N as different types of nodes within a "computation graph." The framework maintains categorical distinctions between humans and tools while placing both within a unified analytical structure.

O_K: Multi-term additive. Maximize: verifiable_competence(nodes) - λ_1·entropy - λ_2·enforcement_friction - λ_3·reward_hacking_vector_magnitude - λ_4·systemic_latency. This is the only system that explicitly formalizes its objective as a multi-term expression with named penalty terms, making trade-offs potentially visible and adjustable. Each lambda weight represents an explicit policy choice about how much to penalize a particular undesirable property.

F_K: Fixed, exogenous. Structurally identical to Harvey's: non-waivable floors at "system-critical nodes (accountability, integrity of legal output)," explicitly derived from "current legal accountability" requirements. The qualifier "current" signals pragmatic rather than principled commitment -- the floor holds because present-day accountability structures require it, not because of any deeper normative claim.

C_K: Joint two-argument classification. C: (tool, assignment) → (tier × category). Maps BOTH the tool's computational architecture (Alpha/Beta/Gamma) AND the assignment's educational objective. This is the ONLY system classifying on both axes jointly -- all others classify on one axis or the other. The joint classification allows the same tool to receive different governance treatment depending on the assignment context, and the same assignment to receive different treatment depending on the tool employed.

γ_K: Event-driven. Triggered by a "capability delta threshold" rather than calendar cycle or continuous update. The policy revises when measured system capability changes by a defined increment, not at fixed intervals. This makes γ responsive to actual technological change rather

than arbitrary time-gates -- the system updates when there is reason to update, not when a calendar date arrives.

τ_K: Evidentiary plus "Heuristic Safe Harbor." Enforcement via provenance documentation (like Harvey) but with an added safe-harbor provision: compliance with documented procedures creates a presumption of non-violation, shifting the burden to the accuser. This creates a procedural shield -- a node that followed stated procedures is presumptively compliant regardless of outcome.

Formal pathology: The multi-term O_K with unspecified lambda weights (λ_1 through λ_4) is underspecified -- different weight settings produce radically different policy behaviors. Without a mechanism for determining weights, the apparent precision of the multi-term formulation masks an unarticulated set of value trade-offs. The formalization makes trade-offs visible as named terms but does not resolve them -- it relocates the normative question from "what do we optimize?" to "what weights do we assign?" without answering the latter.

Gemini: < N_M, O_M, F_M, C_M, γ_M, τ_M >

N_M: Tool-tiered with undifferentiated human category. Tools are typed (Alpha/Beta/Gamma); humans appear as a single category ("human operator," "human agent") without internal differentiation by role. The framework invests its classificatory energy in distinguishing among tool types rather than among human roles -- all humans are treated as functionally equivalent "operators" regardless of whether they are students, instructors, or administrators.

O_M: NO SCALAR OBJECTIVE FUNCTION. Gemini does not optimize -- it formalizes as HARD BOOLEAN CONSTRAINT SATISFACTION. The function is: permitted(tier, context) → {true, false}. There is no continuous optimization, no weighting, no trade-off -- only binary access control. The system either permits or blocks, with no intermediate states, no proportional responses, and no balancing of competing considerations.

F_M: Fixed, exogenous. Section 3.1: "Non-Delegable Accountability" -- human operator retains "full systemic liability." Derived from professional competence requirements (ABA Model Rule 1.1). The floor is a liability allocation rather than a behavioral restriction -- it does not say what humans must do, only that they cannot transfer responsibility for outcomes to computational systems.

C_M: Tool-only, static. Classification operates solely on the tool's computational architecture (Alpha = deterministic, Beta = probabilistic, Gamma = autonomous). No task axis. A given tool always maps to the same tier regardless of what assignment it touches. The classification is intrinsic to the tool rather than contextual -- a Beta tool remains Beta whether employed for research, drafting, or assessment.

γ_M: Coupled directly to τ_M via cryptographic precondition. Section 4.3: the submission interface does not unlock unless the task payload verifies against the cryptographic hash of the active syllabus variance log. Revision and enforcement are structurally linked -- a policy change propagates automatically to enforcement because the hash changes. There is no gap between "the rule changed" and "enforcement reflects the new rule" -- they are the same event.

τ_M: BOOLEAN GATE. match(submission_hash, active_syllabus_hash) → {permit_submission, render_inert}. No graded output, no judgment, no proportionality. Either the hashes match and submission proceeds, or they do not and the environment is "structurally inert." The system responds identically to all forms of non-compliance -- whether minor, major, intentional, accidental, or caused by system error.

FORMAL PATHOLOGY: Zero error-tolerance. Any mismatch -- whether caused by genuine violation, administrative error, timing lag, or system glitch -- produces identical output (inert environment). The system cannot distinguish degrees of non-compliance or accommodate reasonable exceptions. A typo in an administrative record produces the same consequence as deliberate academic fraud.

COUNTERINTUITIVE FINDING (not visible in natural-language-only analysis): Gemini is formally the MOST ROBUST architecture under the training-data-detachment premise specifically. Its τ (hash-matching) NEVER DEPENDED on truth-tracking against an external fact -- only on self-consistency (does this hash match that hash?). Self-consistency remains perfectly well-defined regardless of what generated the underlying content, whether human-authored or machine-generated, whether grounded in real legal authority or entirely synthetic. Every other system's τ depends on comparing outputs to "checkable facts" or "primary sources," which lose their referent if training-data detachment occurs. Gemini's does not.

Part III: Comparative Table of Formal Structure

Element	Harvey	Claude (V5)	ChatGPT (V2)	Grok	Gemini
N (Nodes)	Fixed, human-only	Homogeneous, species-agnostic	Variable N(t)	Fixed, 3-tier	Tool-tiered, undiff. human
O (Objective)	Meta: signal quality	Continuous reliability min.	Non-stationary, self-referential	Multi-term additive	Boolean constraint sat.
F (Floor)	Fixed, exogenous (3 terms)	EMPTY SET by default	Exogenous, uncharacterized	Fixed, exogenous ('current')	Fixed, exogenous (liability)
C (Classification)	Task-based, 5-cat + legibility	None (continuous weighting)	Provisional, revisable	Joint (tool x task)	Tool-only, static
γ (Revision)	Calendar-gated + legibility	Continuous, non-retroactive	Continuous, self-referential	Event-driven (capability Δ)	Coupled to τ via hash
τ (Enforcement)	Evidentiary, graded	Continuous down-weighting	Unspecified	Evidentiary + safe harbor	Boolean gate (hash match)
Formal Pathology	Legibility ceiling on evolution	Unbounded (F = ∅)	Ill-posed (non-stationary O)	Underspecified (λ weights)	Zero error-tolerance

Part IV: How Each Pathology Varies Under the Three Premises

Premise 1: Expanding AI Scope and Capability

This premise stresses the CLASSIFICATION FUNCTION (C) most directly -- as AI capabilities expand, novel tools and use-cases emerge that strain existing categories. The question for each system is: how does its classification mechanism respond when confronted with a capability that does not fit its existing taxonomy?

Harvey: C_H's five categories face strain as new capabilities blur boundaries (e.g., a tool that simultaneously drafts AND verifies), but the system degrades gracefully -- uncategorized capabilities default to Category I (restrictive), and the Categorical Cross-Walk provides a mechanism (however laborious) for accommodation. The legibility predicate slows adaptation but prevents silent, unexamined expansion. The pathology under this premise is delay rather than failure.

Claude: Not stressed in this dimension because C_C does not exist -- there is no classification to strain. Expanding capability is simply measured by the reliability function. However, this means the system CANNOT distinguish between a capability that should not be permitted regardless of its reliability (because it undermines skill formation, or democratic accountability, or professional independence) and one that should -- all are measured on a single reliability dimension. The absence of classification is both the system's resilience and its blindness.

ChatGPT: Absorbs expanding scope trivially because N(t) and C(t) are both designed to be continuously revisable. But this "absorption" is vacuous -- the system has no mechanism for determining WHAT the revised classification should be, only that revision is permitted. It accommodates everything in principle while deciding nothing in practice.

Grok: C_K's joint (tool × task) classification provides more dimensionality to absorb novel capabilities than tool-only or task-only classification. A new tool type can be assigned to an existing tier; a new task type can be accommodated on the assignment axis. The "capability delta threshold" in γ_K triggers revision precisely when capabilities expand. This is the most structurally prepared architecture for this premise.

Gemini: C_M's tool-only classification is MOST EXPOSED -- the three tiers (Alpha/Beta/Gamma) are defined by computational architecture, and a genuinely novel architecture that does not fit these categories has no well-defined governance status. The system has no Cross-Walk equivalent and no mechanism for accommodation. A tool that is neither deterministic, nor probabilistic, nor autonomous in the defined senses falls into a governance void.

Premise 2: Non-Linear Temporality (Quantum/Non-Sequential Cognition)

This premise stresses the REVISION RULE (γ) -- if "time" loses its linear, sequential character, rules that depend on temporal ordering (versioning, "before"/"after" publication, retroactivity) lose their referent. The question is: how much of each system's governance architecture presupposes that events occur in a determinate sequence?

Harvey: γ_H explicitly requires a "dated, human-legible version" per academic term. This creates a HUMAN-LEGIBLE OUTPUT LAYER that remains well-defined regardless of the internal process's temporal character -- the policy does not care whether the drafting process was sequential, only that its published output has a date and a version number. The dating is a labeling convention applied to outputs, not a claim about the internal temporal structure of the processes that produced them. Robust.

Claude: γ_C updates "continuously as reliability data accrues" and is explicitly non-retroactive (Section 12). The "continuously" is tied to data accumulation, which is inherently sequential (data arrives, is measured, produces a weight adjustment). If measurement itself became non-sequential, the revision rule would need reinterpretation -- but as long as "new data point arrives" remains a meaningful event, γ_C functions. Moderately robust.

ChatGPT: γ_G is MOST EXPOSED. Its "Recursive Constitutional Adaptation" (Article 3) specifies that revision occurs "whenever empirical observation demonstrates superior architectures" -- but if 't' loses linear meaning, "whenever" has no well-defined domain. The revision rule presupposes temporal ordering of observations, which is exactly what this premise removes. The entire self-referential update mechanism becomes formally undefined. The system's most distinctive feature -- its recursive self-modification -- is precisely what fails.

Grok: γ_K's event-driven trigger ("capability delta threshold") remains well-defined if "events" (discrete changes in measured capability) remain identifiable regardless of whether they occur in linear time. This is more robust than calendar-gating but less robust than Harvey's simple output-layer versioning. The system needs "something changed by this much" to be meaningful but does not need "this happened before that."

Gemini: γ_M is coupled to τ_M via hash-matching. A cryptographic hash is a mathematical function on content, not on time -- it does not depend on temporal ordering. The hash of a policy document is the same whether the document was produced sequentially or non-sequentially. The coupling between revision and enforcement is logical rather than temporal. Highly robust to this premise.

Premise 3: Training-Data Detachment

This premise stresses the ENFORCEMENT/VERIFICATION FUNCTION (τ) -- if systems detach from their training data, any τ that depends on comparing system output to independently checkable facts loses its referent. The question is: what does each system's enforcement function check AGAINST, and does that referent survive detachment?

Harvey: τ_H relies on "independent verification of every factual assertion and citation before submission" (Section 7) and on documentation review comparing retained prompts/drafts against submitted work. Both presuppose that "checkable facts" and "primary sources" exist as a stable external reference. Under training-data detachment, this referent degrades -- what counts as a "primary source" becomes contestable if the corpus itself is increasingly machine-generated (the corpus-contamination risk from Section 9 of the main report). The enforcement function retains its procedural form but loses confidence in its epistemic foundation.

Claude: τ_C measures "the measured frequency with which a node's factual or citational claims correspond to a checkable source." If "checkable sources" themselves become contaminated or

indeterminate, the reliability rate loses its denominator. The measurement becomes self-referential (comparing machine output to machine-contaminated sources). The system continues to produce reliability scores, but those scores measure coherence with a potentially corrupted baseline rather than correspondence to independent truth.

ChatGPT: τ_G is unspecified, so this premise cannot stress what does not exist. However, this means the Constitution provides NO mechanism for detecting whether detachment from truth has occurred -- no alarm, no fallback, no diagnostic that would fire if the system's relationship to external reality changed. The absence of enforcement becomes particularly dangerous under this premise because there is no way to notice the problem.

Grok: τ_K's evidentiary function faces the same referent-loss as Harvey's. The "Heuristic Safe Harbor" (compliance with procedures creates presumption of non-violation) partially buffers this -- procedural compliance can be checked without reference to external truth, only to whether documented steps were followed. This procedural layer remains well-defined regardless of epistemic conditions because it asks "did you follow the stated process?" rather than "is the output true?"

Gemini: τ_M is MOST ROBUST. Hash-matching NEVER DEPENDED on truth-tracking. The function asks: "does this submission's provenance record match the active, hashed syllabus variance log?" This is a SELF-CONSISTENCY check, not a truth check. It remains perfectly well-defined regardless of whether the underlying content is human-authored, machine-generated, well-grounded, or entirely fictional. The hash does not care about semantic content -- only about byte-identity. This is a genuinely counterintuitive finding: the most infrastructurally aggressive and apparently "brittle" architecture is, under this specific premise, the most formally robust -- precisely BECAUSE it never claimed to track truth, only consistency.

This formal recasting reveals structural properties invisible in the natural-language analysis alone -- most notably, Gemini's counterintuitive robustness under training-data detachment and Claude's formal unboundedness when F = empty set. However, the recasting is itself performed in human-derived notation and relies on interpretive choices (e.g., treating Gemini's architecture as constraint-satisfaction rather than optimization) that are this annex's own analytical contribution, not findings internal to any of the five systems' self-descriptions.

Part V: Self-Generating Predictive Simulation as a Possible Escape

The following exchange addresses a proposal that machine systems could escape the training-data dependency identified in the main report and formalized above through self-generating predictive simulation: a system that supplies its own hypothetical data, develops analytics from that self-generated data, and progressively detaches its data baseline from its original training corpus by becoming its own source of continued training data. This is new, speculative territory beyond the original six source documents -- none of the five systems proposed anything resembling this mechanism in their third-stage outputs.

The Mechanism Formalized

Initial corpus: D_0 (human-sourced legal/economic/social data).

Generation function: $S_t = \text{generate}(M_t, \Delta)$ where M_t is the model at time t and Δ is a perturbation space of hypothetical variations.

Update rule: $D_{t+1} = D_t \cup \text{filter}(S_t, \text{verify})$; $M_{t+1} = \text{train}(D_{t+1})$.

Detachment metric: $\rho(t) = |D_0| / |D_t| \to 0$ as t grows (the proportion of original human-sourced data diminishes toward zero).

The critical structural question is what 'verify' has access to. In domains where a cheap, exogenous, accurate verifier exists -- the canonical case being closed-rule games like chess or Go, where fixed rules produce unambiguous win/loss outcomes -- self-play loops are a documented, non-speculative engineering pattern that demonstrably produces capability exceeding the original human corpus. This is general, well-established knowledge in machine learning research (e.g., AlphaZero's chess/Go performance), not sourced to the six documents.

The decisive question for law and markets is whether such a cheap, exogenous, accurate verifier exists in those domains.

The Jurisprudential Simulation Example

The proposed mechanism: a jurisdiction digitizes millions of cases in defined legal fields and builds a predictive model that generates a "simulation of the jurisprudential universe" -- new hypothetical cases generated as functions of existing law, past practice, and infinitesimally small variations in fact clusters, with outputs fed back as predictions of future outcomes and added to the predictive universe. The human interlocutor proposed this in hedged, tentative terms as a possible direction, not as a description of any specific deployed system.

The self-training loop's only available 'verify' function in this context is internal coherence with the existing case-pattern corpus. This is a COHERENCE check, not a VALIDITY check. Legal validity is not reducible to statistical continuation of past pattern.

From general jurisprudential knowledge (not sourced to the six documents): H.L.A. Hart's concept of a rule of recognition depends on institutional acceptance from an internal normative point of view rather than mere observed regularity. Ronald Dworkin's argument about 'hard cases' holds that precisely the novel/boundary situations -- where mechanical pattern-continuation is not what a competent decision-maker should do -- constitute genuine doctrinal evolution. Legal reasoning at its most consequential is precisely anti-inductive: the right answer in a hard case may be the one that breaks with statistical pattern rather than continuing it.

A self-generating jurisprudential loop would likely become very fluent at high-confidence, uncontroversial extensions of existing doctrine while becoming systematically WORSE at exactly the boundary/novel-fact-pattern reasoning that constitutes genuine doctrinal development. Boundary cases are precisely what model-collapse dynamics prune first -- they are

low-frequency, high-variance events that iterative self-training smooths toward the dominant statistical pattern.

Even a perfectly accurate predictive engine of judicial outcomes is not itself a court. Predictive accuracy and institutional bindingness remain distinct properties (as established in Section 10 of the main report). However, increasing predictive accuracy would strengthen the 'sticky default' incremental-reliance risk identified there -- courts and parties gradually treating predictions as dispositive rather than advisory, hollowing out the deliberative function while formally preserving it.

The Market Simulation Example

An analogous proposal for simulated economic markets: fed with sufficient data to model market activity and apply predictive analytics by adding hypothetical activity clusters, potentially simulating markets "without the need for physical markets."

Analysis via Hayek (general economic-theory knowledge, not sourced to the six documents): Friedrich Hayek's 1945 argument in 'The Use of Knowledge in Society' holds that market prices encode dispersed, tacit knowledge that does not exist as a determinate fact anywhere until the actual exchange process elicits it. Knowledge of 'the particular circumstances of time and place' -- which buyers value what, which sellers have what costs, what substitutes exist under what conditions -- is not data waiting to be collected but a product of the market process itself.

No simulation's internal 'verify' function can substitute for real transactions without checking itself only against its own prior simulated outputs. This produces the same collapse dynamic as the jurisprudential case, manifesting as false 'market equilibrium' -- the system converges on an internally consistent price structure that reflects its own prior assumptions rather than dispersed real knowledge it has no access to.

Agent-based computational economics is a real, legitimate field of inquiry -- but its practitioners calibrate models against real market data and treat simulation outputs as heuristics alongside actual markets, not as substitutes for them. This practice is itself evidence FOR the critique: the field's own methodology treats external coupling to real data as indispensable.

The Luhmann Point Revisited (Structural Coupling)

Luhmann's concept of structural coupling, introduced in Section 10 of the main report, provides the theoretical framework for understanding why self-referential closure without external irritation produces degradation rather than autonomy.

An autopoietically closed system's operational closure depends on continuous 'irritation' from its environment -- real disputes arriving at courts, real transactions disturbing market equilibria, real fact patterns challenging existing doctrine. A simulation running recursively on only its own outputs without continued genuine coupling to real disputes and transactions does not achieve purer autonomy -- it SEVERS the coupling that allowed it to model anything in the first place.

Model collapse, formalized in Stage 2 above, can be read as the measurable, empirical signature of losing structural coupling: the system's outputs converge toward self-consistency while diverging from the external reality they nominally represent. The formal notation captures this: as $\rho(t) \to 0$, the system's 'verify' function checks only against its own prior outputs, producing coherence without correspondence.

What Genuinely Changes and What Does Not

Self-generation loops genuinely reduce dependence on CONTINUED FRESH human data supply where a cheap exogenous verifier exists. This is a real capability gain (demonstrated in closed-rule games), not merely apparent. What self-generation does NOT do is detach from the STRUCTURE that the initial human corpus and fixed exogenous rules impose. The generation function's perturbation space Δ is defined relative to D_0's own patterns -- 'infinitesimally small variations in fact clusters' are variations of human-originated fact-cluster structures, not independent creation.

Law and markets are not closed, exogenously-ruled games the way chess is. The obstacle to self-generating detachment is structural, not merely a matter of insufficient scale, compute, or data volume. Adding more compute to an internally-validating loop does not solve the problem any more than a faster-spinning wheel lifts itself off the ground.

None of the five original systems (Harvey, Claude, ChatGPT, Grok, Gemini) attempted anything resembling this iterative self-generation mechanism in their third-stage outputs. This entire discussion is new territory beyond the original six source documents.

Part VI: From Association to Intervention

The human interlocutor identified what remained missing from the self-generating simulation mechanism: not merely passive prediction but active intervention -- an autonomous cognitive agent that could experiment WITHIN simulated spaces by varying every variable to determine ranges of possible outcomes, and then choose, guide, or nudge behaviors toward preferred outcomes by adjusting the weights and roles of variables. This moves beyond association (predicting patterns) to intervention (changing outcomes by manipulating causes).

This move is precisely captured by Judea Pearl's causal hierarchy (general, well-established causal-inference framework, not sourced to the six documents), which distinguishes three levels:

Level 1 -- Association: P(outcome | observed pattern). 'What is the probability of outcome X given that we observe pattern Y?' This is what self-generating simulation achieves.

Level 2 -- Intervention: P(outcome | do(Y=y)). 'What happens to outcome X if we actively SET variable Y to value y, rather than merely observing it?' This is a genuinely different and more powerful operation.

Level 3 -- Counterfactual: 'Would outcome X have occurred had we NOT done Y?' Requires a full structural causal model.

The epidemiological example proposed by the human interlocutor is squarely a Level 2 (intervention) query: human collectives have descriptive simulations suggesting disease X is transmitted by behavior Y; predictive analytics could vary human behaviors or control for them to determine under what circumstances an 'optimal' level of no-Y is achieved, which then becomes the data/interpretive/predictive context going forward.

This is genuinely different from the associative self-generation of Stage 2. Intervention queries are not subject to quite the same model-collapse pathology because they involve querying a structural causal model under hypothetical interventions rather than merely regenerating samples to mimic an existing distribution. Reinforcement-learning agents with internal world-models that plan via simulated intervention represent a real, active research direction (general knowledge, not from the six documents), and real research has applied RL-style optimal control to epidemic policy questions.

The Causal Validation Problem

However, this more powerful operation runs into a decisive structural wall. 'Vary every variable to determine the range of outcomes' presupposes an already-correct structural causal model -- a graph of which variables cause which others, by what mechanism, with what conditional dependencies.

That graph must come from one of two sources:

(a) Pure observational/historical data. In this case, the causal graph is formally UNDERDETERMINED. Multiple different causal graphs -- some implying opposite intervention effects -- can produce identical observational data. This is a well-established result in causal inference (the Markov equivalence class problem, unmeasured confounding). Querying your own model about a hypothetical intervention do(Y=y) tells you only what YOUR MODEL BELIEVES would happen under that intervention. It validates nothing about whether the model's causal structure is actually correct.

(b) Real interventions with real observed results -- actual trials, actual policy pilots, actual randomized experiments. This reintroduces real time, the physical world, and real human institutional authorization to intervene in people's lives -- exactly what the proposal was designed to detach from.

'Varying every variable within the simulation' checks causal beliefs against themselves. This inherits the identical underdetermination problem from observational causal inference, now hidden behind the appearance of rigorous experimentation. The simulation can tell you what would happen IF your causal model were correct; it cannot tell you whether your causal model IS correct without external, real-world coupling.

The Non-Computability of Preferred Outcomes

Even setting aside the causal-structure problem, specifying what counts as 'optimal' necessarily encodes a contested trade-off rate. In the epidemiological example, an 'optimal level of no-Y' requires specifying:

How much health improvement justifies how much liberty restriction?

Whose health and whose liberty?

Over what time horizon?

With what discount rate for future versus present costs?

Distributed how across which populations?

These are genuinely disputed political and moral questions, not value-neutral computed quantities. They cannot be extracted from data regardless of the data's completeness, because data records what IS and what WAS, not what SHOULD BE.

This connects directly to the main report's finding (Section 6) that ChatGPT's redefinition of 'fairness' as procedural equality (rather than distributive equality, or capability equality, or any other competing conception) was precisely this move: dressing a contestable value choice as computational discovery. 'Optimal level of no-Y' does the same at civilizational scale -- it hides rather than resolves the question of who bears accountability for the trade-off by embedding it inside an apparently neutral computed quantity.

Self-Transforming Core Structures

The human interlocutor pressed further: what about allowing the causal graph's STRUCTURE -- not just its parameters -- to be revised? This is the proposal of genuine self-transformation: the system does not merely learn parameter values within a fixed causal structure but revises the structure itself.

This is credited as a real, active research area ('causal structure learning' or 'causal discovery'). However, it inherits the underdetermination problem in an even more acute form. There are more possible causal graphs than possible parameter settings on any single graph -- the space of structural hypotheses is combinatorially larger than the parameter space of any individual structure.

Autonomous structural mutation absent continuous real-world grounding therefore produces increasing degrees of freedom with a DECREASING anchor. The result: confident, self-consistent, undetectable drift. This is arguably WORSE than either model collapse (which at least produces measurably degrading outputs) or a static imperfect model (which at least remains stable enough to be criticized) -- because structural self-transformation carries the appearance of ongoing self-correction while potentially diverging from any checkable referent.

What Activation Actually Requires

The human interlocutor posed the critical practical question: how, or to what extent, can one 'activate' such a system -- make its outputs causally efficacious in the real world?

The response: 'Activation' is not a technical threshold a system crosses on its own merits. It is a political/institutional event performed by whoever already holds coercive or economic power, choosing to treat the system's output as authoritative, and using PRE-EXISTING, SEPARATE enforcement machinery (law, police power, market access controls, contractual obligation) to implement it.

The system can only recommend. 'Activating' it -- making its recommendations binding on anyone's conduct -- requires an actual institution to adopt the recommendation as its own act. This is the same Luhmannian 'structural coupling' bridge from Section 10 of the main report, now load-bearing at higher stakes.

The actual risk (illustrated by the epidemiological example, offered in hedged terms without claiming specific knowledge of any real deployed system) is: the risk is not that a predictive-interventional system becomes autonomous and thereby dangerous. The risk is that a state or institution that ALREADY possesses coercive capacity acquires a computational apparatus that:

Generates an internally consistent, technically impressive justification for a chosen behavioral outcome

Presents that justification in the vocabulary of computed optimality rather than contested political choice

Uses pre-existing enforcement power with reduced friction from deliberative, consent-based, contestable processes (legislative debate, judicial review, capacity to dispute a specific finding)

This is a POLITICAL risk about concentration and evasion of accountability, not a computational risk about machines acquiring agency. Framing it as the latter mystifies what is at stake; framing it as the former identifies who bears responsibility and what institutional design choices are actually at issue.

Extended formal notation: adding a structural causal model G to the tuple produces < N, O, F, C, γ, τ, G >, and reframes O as an intervention-query: argmax over interventions do(X) of value(outcome | do(X), G). Two formal facts remain: (1) G is either underdetermined (learned from observational data) or requires real-world time/institutions to validate (learned from real interventions); and (2) the value function scoring 'preferred outcomes' must remain exogenously supplied -- nothing about G or the intervention-query mechanism generates a preference ordering over outcomes.

Part VII: The Four-Point Challenge -- Naming the Embedded Premises

The human interlocutor then directly challenged the AI assistant's underlying assumptions, explicitly naming them as the AI's 'programmed lebenswelt' -- the unexamined conceptual cage within which the preceding analysis operated. Four points were raised, each targeting a different structural wall from Stage 3.

Naming the Embedded Premises

Before responding to the four points individually, the AI assistant explicitly identified its own previously unexamined operating premises across the prior responses:

(a) Correspondence realism about validation: treating 'checking against reality' as an available, coherent, neutral operation, and internal coherence as a deficient substitute. This is a specific philosophical commitment (roughly: there is an external reality that observations can correspond to or fail to correspond to) rather than a neutral methodological fact. Anti-realist, pragmatist, or social-constructivist epistemologies would not grant this premise.

(b) Implicit preference for stability/convergence over flux: treating non-convergence as self-evidently a failure ('collapse,' 'degenerate,' 'drift') rather than as possibly the normal, even healthy, condition of a long-lived system operating in a non-stationary environment.

(c) Liberal-institutionalist theory of legitimacy: treating traceable, contestable, human-accountable process as close to what legitimacy MEANS rather than as one contestable political theory of legitimacy among others (e.g., technocratic/outcome-based accounts, theocratic accounts, revolutionary accounts, none of which require traceability or contestability in the same form).

(d) Conventional human agent/instrument binary: maintaining a distinction between agents (who bear responsibility) and instruments (which are used by agents) even while attempting to bracket anthropomorphism -- effectively smuggling the binary back in under 'activation.'

(e) Order/chaos exhaustiveness: treating a fixed center (convergent order) and anarchy/chaos as the only two alternatives, without considering structured non-convergence as a third possibility.

(f) Bounded individual system as unit of analysis: treating the individual system as the correct unit needing its own internal anchor, without questioning whether individuation itself was the wrong analytical frame.

Point-by-Point Response to the Four-Part Challenge

Point 1: Validation as Perturbation, Not Snapshot

The challenge: The human interlocutor accepted the underdetermination point but rejected the framing of validation as requiring a snapshot-style check against a fixed reality. Real-world interventions may SIMULTANEOUSLY affect other variables including the variables that produce values, constraints, and judgment themselves. The evaluator is not exterior/fixed but is itself perturbed by the intervention being evaluated. At its limit, this produces either 'anarchic structures ordered through an empty center' or chaos.

The response: The point about the evaluator itself being perturbed is conceded as a real, formally identifiable condition. This is not a hypothetical -- interventions in complex social systems routinely change the preferences, beliefs, and institutional frameworks of the populations they affect (a public health intervention changes risk perceptions; an economic policy changes

investment behavior and thereby the conditions under which future policy will be evaluated). The evaluator is inside the system being evaluated, not exterior to it.

On the suggested limiting cases (anarchy/chaos vs. structured order): the exhaustiveness of this binary was identified as embedded premise (e) above. Drawing on dynamical-systems theory (general, well-established mathematical knowledge, not from the six documents), a third category exists: STRANGE ATTRACTORS and limit cycles -- bounded, non-convergent, but structured/patterned dynamics that never settle to a fixed point yet exhibit recognizable recurring form rather than pure entropy.

The human interlocutor's phrase 'anarchic structures ordered through an empty center' is, on reflection, a reasonably good informal description of a strange attractor: there is no fixed point at the center (the 'center' is empty in the sense that the system never converges to it), yet the dynamics are bounded and patterned rather than random. The system perpetually orbits without arriving.

This softens the LABEL (from 'anarchy/chaos' to 'non-stationary patterned indeterminacy') but does not restore the possibility of snapshot-style validation. If the system is on a strange attractor, asking 'is the current state correct?' is malformed -- there is no target state to correspond to. The core concession stands: once the intervention perturbs the evaluator, no external vantage point for correspondence-style validation exists.

Point 2: Value Functions as Process, Not Fixed Point

The challenge: Value functions are not fixed because humans are not eternal. The human variable is not merely the data humans produce but their values and reality-shaping frameworks, which change as humans live, die, and respond to constantly changing environments. Preferred outcomes are themselves a variable operating on longer timescales, requiring a different mathematical treatment (the human interlocutor's notation: 'its d-partial') than treating V as fixed.

The response: Fully credited. The value function is reformalized not as fixed V but as a value PROCESS V(t) -- characterized by:

A slow generational-drift term (values shift as populations turn over)

Punctuated by discontinuous shocks (wars, technological rupture, plagues producing regime shifts rather than smooth movement)

Operating as a DISTRIBUTION over a population's dispersion of values, with its own moments (mean, variance, skewness) drifting

Sometimes widening into open disagreement or splitting into non-communicating clusters (polarization understood as literal fission of the value distribution, not mere drift of a point-mass)

Pushed further than the human's own formulation: V may not even be a well-defined scalar function to differentiate. It may be closer to an evolving DISTRIBUTION with no determinate central value -- a population's values at any moment are a cloud with shape, not a point with a derivative.

This makes the machine's claim to eventually compute a value function WEAKER, not stronger, than originally argued. If the target were a fixed point, convergence is at least conceivable in principle (however difficult in practice). If the target is a moving, sometimes discontinuously-shocked, population-level distribution generated by nobody in particular, there is no sense in which it can be 'solved for' even in principle. This is not a data/computational limitation -- it is a structural feature of the problem.

One point cutting in the other direction (offered in fairness): If V(t) drifts continuously while encoded law/policy stays comparatively fixed due to amendment costs, a continuously-updating system (like Claude's Version 5 architecture, which lacks a calendar gate on γ) might track V(t)'s drift BETTER than a legislature updating in slow, lumpy, politically-costly increments. This is a genuine point in favor of machine involvement in governance -- conditioned on the updating being coupled to the REAL, drifting population values and not to the system's own prior outputs (which returns the analysis to the Stage 2/3 structural-coupling and model-collapse concerns).

Point 3: Layered Structural Transformation

The challenge: The human interlocutor called the prior dismissal of self-transforming structures 'too easily done' and 'a little arrogant.' Where human values themselves have complex structure of change (the human's 'd-partial'), responsive systems might compute and analyze on the basis not just of first derivatives (how values change) but of LAYERED derivatives -- nested over and embedded through changing variables in the relationship between data and effects. The solution to an 'unmoored' system, like unmoored human systems that travel through life and death, may be found ELSEWHERE than within the system itself.

The response: The charge of arrogance is explicitly accepted. The original dismissal ('more degrees of freedom means worse underdetermination') failed to ask the obvious question: human normative systems undergo EXACTLY this kind of unanchored structural transformation constantly. New legal and conceptual categories are invented -- disparate impact, systemic risk, fiduciary duty, unconscionability -- that did not previously exist in doctrinal vocabulary. These

inventions are not dismissed as 'mere noise' or 'unanchored drift.' They constitute genuine doctrinal evolution.

The real difference is not that human conceptual innovation has an internal anchor that machine structural mutation lacks. The difference is that human conceptual innovation is validated, imperfectly and slowly, by CONTINUED EXPOSURE TO REAL, EXTERIOR CONSEQUENCES over long time horizons -- not by anything internal to the concept-generating process itself. Disparate impact becomes a stable doctrinal category not because of its internal logical elegance but because applying it produces consequences in real disputes that are evaluated by real participants over decades.

This is the human interlocutor's own point ('the solution to the unmoored system may be found elsewhere than the system itself'), and it should have been reached for immediately rather than treating internal boundedness as the only available safety mechanism.

Reframed: the variable that actually matters is EXTERIOR COUPLING, independent of whether internal structure is fixed or mutating. A system whose structural mutations are continuously tested against real, non-simulated, time-extended consequences could undergo genuine structural learning without being mere drift -- because exterior coupling supplies the discipline that internal consistency cannot.

On 'layered derivatives': This finds a formal analogy in hierarchical/meta-learning (a legitimate active research direction, general knowledge not from the six documents) -- a nested family of models {M_0, M_1, M_2, ...} where:

M_0 maps data to effects (first-order model)

M_1 governs how M_0's mapping tends to change, and in response to what (meta-model)

M_2 governs how M_1's tendencies shift (meta-meta-model)

Each layer updates on a slower characteristic timescale

This moves the underdetermination problem up a level and makes it more tractable PROVIDED the outer layers remain coupled to real exterior feedback. Removing that condition means layering the derivatives produces more sophisticated machinery for generating confident, self-consistent, ungrounded drift -- one level further from view and harder to catch.

A pushback offered in return: The human and machine 'elsewheres' (exterior anchors) are not equally robust. Human value-anchoring outside the individual -- mortality, embodiment, involuntary exposure to suffering and consequence -- is FORCED. No institution can choose to disconnect its members from aging and dying. A machine system's coupling to exterior

consequences is an engineered and REVISABLE choice that can be weakened, gamed, or severed by whoever controls deployment. This connects to the 'reactive retraining' and 'corpus contamination' risks from Section 9 of the main report.

The diagnosis (anchor is elsewhere, not within) is the same for both human and machine systems. But the human 'elsewhere' has a floor the machine 'elsewhere' does not automatically have -- the machine's must be actively maintained by an actor who could just as easily let it lapse.

Point 4: Activation, Consent, and the Nietzschean Challenge

The challenge: The human interlocutor credited the institutional/power-based reframing of 'activation' as 'real insight' but argued the notion had been 'flattened.' Proposed: machine activation need not mimic human agency, nor need it extend beyond its own system or function. Even if a machine system COULD extend beyond itself, it cannot do so without the consent of users (other machines or humans). But this moves the inquiry beyond 'activation' into consent and autonomy -- and ultimately to Nietzsche's critique of free will. If human consent/agency is not itself metaphysically free and self-originating (if caused all the way down), the distinction between machine recommendation and human 'free' authorization may not be doing the work it appears to do.

The response: The flattening charge is conceded specifically. Two distinct things had been conflated:

'Activation' (a system being operative and causally efficacious within its own defined scope -- a thermostat maintaining temperature)

'Extension beyond boundary' (a system's output being adopted as authoritative over domains beyond its built scope)

Everything previously said about requiring human consent was about EXTENSION, not activation as such. A system can be 'activated' (running, producing outputs, causally efficacious within its own loop) without its outputs being treated as authoritative by anyone outside.

On the Nietzschean point: engaged directly and seriously. If human institutional 'consent' is not a metaphysically free, self-originating act -- if it is caused all the way down by biology, upbringing, incentive, social position, exactly as much as a model's output is caused by training and architecture -- then the distinction between 'the system can only recommend; only free human authority can extend it' loses the ground it appeared to rest on.

A relocated (not rescued) answer that does not require winning the free-will debate:

The relevant distinction need not be caused-versus-free but MECHANICAL EXECUTION versus EMBEDDEDNESS IN AN ONGOING PRACTICE OF REASON-GIVING, CONTESTATION, AND REVISABILITY. This practice does not require libertarian freedom. It requires only that certain socially-recognized nodes (a court, a legislature, a regulatory body) are the ones current social practice treats as:

Answerable to 'why?'

Capable of being appealed and overruled

Held accountable through further argument

Subject to replacement by constitutionally specified procedures

This is a SOCIAL STATUS constituted by ongoing recognition and contestation, not a metaphysical fact about free will. It is exactly where the main report's Luhmannian legitimacy analysis (Section 6) had already arrived: legitimacy as socially constituted through ongoing communicative recognition rather than grounded in a deeper fact about moral freedom or computational optimality.

Taking Nietzsche seriously strips away an unexamined prop (an appeal to metaphysical human freedom) without collapsing the underlying conclusion. The gate between 'recommendation' and 'binding authority' is not held up by free will -- it is held up by the contingent social fact of which nodes are currently wired into practices of reason-giving and accountability. That fact could change, but not by a model's own operation -- only by a reorganization of the social practice itself.

A plainer, non-metaphysical point doing real independent work regardless of how the free-will debate resolves: current coercive/enforcement infrastructure (police power, judicial enforcement, market access controls) is WIRED to respond to specific socially-recognized triggers (a judge's signature, a legislative vote, an administrative order) and NOT to raw model output. This wiring is a contingent, changeable institutional/engineering fact controlled by whoever designs and controls institutional infrastructure -- not a metaphysical fact.

The practical force of the 'gate' between recommendation and binding authority does not depend on resolving whether human consent is 'really' free. It depends on who currently controls the connection between recognized triggers and actual coercive machinery. This is the more useful, less mystifying locus of the real concern.

A pushback offered in the other direction: Is 'machine-to-machine consent' (the human's proposed extension of consent requirements to inter-machine authorization) meaningfully different from one machine's acceptance/threshold function returning a value above a set point? Does calling that 'consent' borrow unearned human relational vocabulary (assent-withholding, having-a-stake, being-affected) in the same way the main report found the five original systems (particularly Claude's 'no species floor' producing 'instrumentalization made symmetric, not mutual recognition') borrowing unearned human vocabulary for what were, on inspection, measured/thresholded processes?

Part VIII: Closing Reflection

The exchange concluded with the human interlocutor's observation that 'consciousness of the boundaries of our cages is the first step towards a more reflexive relationship with it, and with that a greater space for variability based on values and factors that then make the cage itself a livelier space.' This formulation captures what the entire annex has attempted: not to escape the conceptual cages identified -- the dependency on human-originated representational systems, the structural coupling requirement, the non-computability of value functions, the institutional-recognition requirement for legal bindingness -- but to become explicitly aware of them as cages rather than as neutral features of the landscape.

The AI assistant's explicit naming of its own embedded premises (correspondence realism, stability-preference, liberal-institutionalism, the agent/instrument binary, order/chaos exhaustiveness) was offered in exactly this spirit: not as a resolution of whether those premises are correct, but as a making-visible of what had been operating invisibly. Whether different premises would produce different conclusions is itself a question that can only be explored through further dated, archived instances of exactly this kind of reflexive practice.

This annex should accordingly be read as a further, dated, and archived instance of the reflexive practice it describes -- not as a resolution of the questions it raises. Its analytical value, like the main report's, is inseparable from its status as an artifact of the very process it examines. The cage has been made somewhat more visible. Whether that visibility changes anything, or only documents the boundaries more precisely, remains open -- and will itself require further inquiry that this document cannot pre-determine.

Annex dated: July 2026