

From Ethical Principles to Governable Action: A Priority Architecture for Context-Sensitive Normative Governance of Large Language Models

Michael Bächle

baechle@dhbw-ravensburg.de

Baden-Wuerttemberg Cooperative State University

Research Article

Keywords: Responsible AI; AI ethics; normative governance; large language models; ethical prioritization; normative traceability

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10961075/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

From Ethical Principles to Governable Action: A Priority Architecture for Context-Sensitive Normative Governance of Large Language Models

Abstract

Responsible artificial intelligence frameworks converge around principles including non-maleficence, beneficence, autonomy, justice, privacy, responsibility, and explicability, but they do not determine action when principles conflict or specify how contextual priorities should be enforced. This paper introduces the *Priority Architecture* (PA), a metamodel for context-sensitive normative governance of large language models (LLMs). PA represents a traceable refinement chain from a disclosed normative foundation through Context Dimensions, Priority Patterns, and Resolution Modes to Action Envelopes and enforcement, while externalizing priority resolution from the model. A proof of concept instantiates PA for an enterprise-internal LLM assistant handling synthetic private contact data and replicates a fixed protocol with GPT-5.6 Sol and Claude Sonnet 5. GPT-5.6-Sol-based coding classified 100/120 baseline responses and 120/120 responses under strong policy prompting and external PA governance as compliant for GPT; for Claude the counts were 92/120, 108/120, and 120/120. The prompt-versus-PA difference was confined to two difficult case types and was measurement-sensitive. The 1,440-response coding was evaluated on a stratified 20% human sample. Two independent human coders agreed on 281/288 items (97.6%); among these consensus items, GPT-5.6 Sol agreed on 275/281 (97.9%). No extraction errors occurred in the benchmark, but ambiguous or adversarial inputs were not tested. The results support the narrower claim that normative priority resolution can be externalized into an explicit, deterministic, machine-readable, and auditable governance object; they establish neither model independence nor inherent ethical superiority.

Keywords: Responsible AI; AI ethics; normative governance; large language models; ethical prioritization; normative traceability

1 Introduction

The contemporary debate on responsible artificial intelligence (AI) is no longer characterized by an absence of ethical principles. Across guidelines, frameworks, and governance initiatives, a recurring vocabulary has emerged around transparency, justice and fairness, non-maleficence, responsibility, privacy, autonomy, and beneficence

(Floridi et al., 2018; Jobin, Ienca, & Vayena, 2019). This convergence is normatively significant because it establishes a shared repertoire for criticizing and governing AI. It is also incomplete. As Jobin et al. (2019) show, convergence in labels coexists with substantive divergence in interpretation and implementation. A principle such as privacy, autonomy, or safety does not by itself specify how a system should act when another ethically relevant principle points in a different direction.

This implementation problem becomes acute for large language models (LLMs). Recent empirical work demonstrates that changing the prioritization of ethical principles within the same ethical framework can change LLM decisions and reduce consistency (Kotzanikolaou, Dasaklis, Christodoulakis, & Malamas, 2026). Related research on alignment also cautions against treating compact slogans such as helpfulness, harmlessness, and honesty as if they settled the underlying normative and political questions of AI governance (Dahlgren Lindström et al., 2025). The problem, therefore, is not merely to state ethical values more clearly. It is to make explicit how normative commitments are refined into action-guiding constraints under context.

Existing work supplies important parts of this transition. Ethical governors separate normative assessment from agent action (Cardoso, Ferrando, Dennis, & Fisher, 2022; Dennis, Fisher, Slavkovik, & Webster, 2016); Rulebooks model partially ordered requirements (Censi et al., 2019); defeasible deontic systems represent obligations, exceptions, and overrides (Neufeld, Bartocci, Ciabattini, & Governatori, 2022); LogiKEy supports formal normative theories (Benzmüller, Parent, & van der Torre, 2020); and GenEth derives explicit principles from expert judgments (Anderson & Anderson, 2014). In autonomous driving, Geisslinger, Poszler, and Lienkamp (2023) combine a non-compensatory risk boundary with weighted ethical objectives and context-dependent safety priority. These approaches show that operationalization is possible, but they emphasize different segments of the path from ethical justification to technical enforcement.

The central claim is that this path should itself be a governable artifact. The proposed *Priority Architecture* (PA) makes explicit the adopted ethical foundation, principle interpretations, admissible context, context-sensitive priority relations, resolution logic, operational governance decision, and resulting action constraints. PA thus addresses the interface between moral justification and technical AI governance (Dignum, 2018; Hammerschmidt, 2025).

PA is not offered as a universal ethics algorithm. It is ethically open at the meta-level and deliberately separates *formalization* from *legitimation*. A priority can be represented and enforced without thereby becoming democratically, legally, professionally, or morally justified. This distinction is essential because technical control over AI does not itself confer legitimate normative authority (Roberts, 2024). The role of PA is narrower but practically consequential: to make an adopted normative path explicit, inspectable, versionable, and technically actionable.

The paper has two linked aims. The conceptual aim is to define PA as a metamodel for context-sensitive normative governance. The empirical aim is to test whether a small instantiation can externalize priority resolution from an LLM in a controlled setting. The central research question is:

How can an explicitly adopted normative position on competing Responsible-AI principles be represented and operationalized as a traceable, context-sensitive refinement chain leading to enforceable governance rules for LLM-based systems?

The proof of concept asks whether such a chain can be represented deterministically and integrated with an LLM without delegating normative priority to the LLM. The paper contributes (1) a normative refinement metamodel linking foundations, admissible context, priority structures, resolution modes, operational governance, and enforcement; (2) *Normative Traceability* and explicit **unresolved** states as governance properties; and (3) a controlled cross-model demonstration that separates Context Acquisition, deterministic Normative Resolution, and downstream LLM compliance so that failures can be localized.

The empirical scope is deliberately narrow: one synthetic enterprise privacy domain, three principles, four Priority Patterns, two implemented Resolution Modes, and two proprietary LLM families. The study tests the *Specification Fidelity* and integration of one concrete PA profile, not its external normative validity, domain transferability, or model independence.

2 Related Work and the Normative-Operationalization Gap

2.1 Principle convergence does not settle implementation

The literature on AI ethics has repeatedly identified convergence at the level of high-level principles. Floridi et al. (2018) synthesize beneficence, non-maleficence, autonomy, and justice and add explicability, understood as combining intelligibility and accountability. Jobin et al. (2019) similarly identify recurring clusters across AI-ethics guidelines while emphasizing divergence in interpretation, scope, justification, and implementation. This divergence matters because governance cannot operate on principle names alone. Fairness concerns, for example, include distributional consequences and the possibility of disproportionate impacts on vulnerable groups (Eom et al., 2026). The operational problem therefore begins before a rule is executed: the meaning of the principle in the application domain must be specified.

Current work likewise argues for context-sensitive, multi-objective ethical frameworks that accommodate different normative perspectives (Hammerschmidt, 2025). For PA, this means that plural considerations should not be silently collapsed into one numerical objective: some requirements may be constraints, priorities, trade-offs, or reasons for human deliberation.

Explainable artificial intelligence (XAI) addresses how model outputs can be understood, with recent reviews emphasizing interpretability, accountability, fairness, and lifecycle governance (Camilleri, 2026). PA concerns a different object: *Normative Traceability* asks why an action was permitted, required, restricted, or escalated under the adopted governance framework. A model may therefore be explainable in an XAI sense while the normative basis of permitted action remains opaque, or normatively traceable while the model itself remains a black box.

2.2 Conflict resolution requires more than one mechanism

Machine ethics offers distinct conflict mechanisms. Weighted optimization permits compensation; lexicographic ordering protects explicit precedence (Wang, Huang, Khajepour, Cao, & Lv, 2020); Rulebooks allow partial orders and incomparability (Censi et al., 2019); Maximin privileges the worst-off; and defeasible deontic logics encode exceptions and rule superiority (Neufeld et al., 2022). Human deliberation remains necessary where the adopted framework intentionally refuses automatic resolution.

The autonomous-driving work of Geisslinger et al. (2023) is especially instructive because it combines several forms of normative reasoning. A maximum acceptable risk functions as a threshold that constrains the feasible set. Within admissible options, overall risk, equality, Rawlsian Maximin, and responsibility are incorporated into a cost function. When no normally valid trajectory is available, the decision logic changes and risk receives priority over mobility and comfort. This design illustrates two principles that PA generalizes: constraints should be evaluated before compensatory optimization, and the appropriate resolution mechanism can depend on context.

This pluralism reflects different logical roles assigned to norms: rights and duties may act as boundaries, consequentialist considerations evaluate outcomes, Maximin privileges the worst-off, proportionality structures balancing, and procedural ethics may require review. Forcing all roles into one weighted sum would embed a substantive moral theory without making that choice visible.

2.3 Formal ethical reasoning and external governance

Machine-readable ethical reasoning provides further building blocks. Formal verification can assess ethical choices (Dennis et al., 2016); defeasible normative supervisors can constrain learned policies (Neufeld et al., 2022); LogiKEy supports normative theory design (Benzmüller et al., 2020); and LegalRuleML represents normative rules with metadata, provenance, and authority information (Palmirani et al., 2021). Together, these approaches support separating inspectable normative structures from adaptive AI components.

External governors similarly separate agent behavior from normative control (Cardoso et al., 2022). GRACE, for example, combines a moral module that determines permissible macro-actions, an optimizing agent, and an enforcement guard (Jahn, Muskalla, Dargasz, Schramowski, & Baum, 2026). This resembles the technical boundary between PA, an Action Envelope, and downstream guardrails.

Case-based approaches address where operational principles come from. GenEth infers principles from judgments over dilemmas (Anderson & Anderson, 2014), and clinical work operationalizes prima-facie ethical principles using clinical cases in a domain-specific decision model (Meier, Hein, Diepold, & Buyx, 2022). For PA, such methods may help identify candidate Priority Patterns, but normative authorization still requires appropriate expertise and institutional legitimacy.

2.4 Policy-based and attribute-based authorization architectures

Policy and access-control architectures provide an important runtime precedent. Attribute-Based Access Control (ABAC) evaluates subject, object, operation, and environmental attributes against explicit policies (Hu et al., 2014). The eXtensible Access Control Markup Language (XACML) separates policy administration, information, decision, and enforcement and supports combining rules and obligations (Rissanen, 2013). A PA profile can therefore resemble established policy architectures at runtime: structured context is evaluated, a decision is produced, and enforcement remains separable.

The distinction lies upstream. Conventional policy mechanisms take policies, attributes, and combining semantics as given. PA additionally represents the adopted normative path: why context is admissible, how principles are interpreted, when Priority Relations apply, which Resolution Mode is authorized, and which authority supports the resulting rule. PA does not generate or legitimate these commitments; it makes them explicit, versionable, and auditable. It is therefore complementary to, not a replacement for, policy-decision and enforcement infrastructures.

2.5 Research gap and design requirements

The approaches above are complementary rather than mutually exclusive. Among the literature reviewed for this study, no single established approach identified spans the full chain from explicitly disclosed normative foundation through context-sensitive priority resolution to operational governance, action constraints, and enforcement. PA is proposed as a metamodel that connects these layers rather than as a replacement for specialized methods at any one layer, including policy-decision and policy-enforcement mechanisms.

Table 1 summarizes the relationship between representative approaches and the PA design problem. The table is interpretive: it identifies the primary contribution of each research stream rather than claiming that a cited approach cannot be extended beyond the indicated function.

Six design requirements follow: **DR1**, disclose the normative foundation; **DR2**, admit only normatively justified Context Dimensions; **DR3**, support plural conflict-resolution modes; **DR4**, preserve normative provenance; **DR5**, expose missing specification as **unresolved**; and **DR6**, represent governance decisions independently of downstream generation.

Table 1 Representative approaches and their relation to the Priority Architecture design problem

Approach	Primary mechanism	Relevance and remaining gap for PA
AI-ethics principles (Floridi et al., 2018; Jobin et al., 2019)	Principle frameworks	Establish normative vocabulary; do not by themselves specify context-sensitive priority resolution or execution
Rulebooks / lexicographic control (Censi et al., 2019; Wang et al., 2020)	Partial or lexical priority	Make precedence explicit and resist unrestricted compensation; normative provenance and full refinement chain remain external
Constraint-first ethical optimization (Geisslinger et al., 2023)	Threshold plus plural optimization	Demonstrates context-dependent changes in resolution logic; highly domain-specific
Defeasible deontic reasoning (Neufeld et al., 2022)	Obligations, permissions, exceptions	Supports explicit rules, overrides, and auditability; does not by itself define the upstream ethical refinement process
Normative meta-frameworks (Benzmüller et al., 2020; Palmirani et al., 2021)	Formal normative representation	Strong basis for machine-readable norm systems and provenance; not specifically an LLM governance pipeline
Expert elicitation (Anderson & Anderson, 2014; Meier et al., 2022)	Case-based principle derivation	Useful for deriving and validating candidate patterns; domain and expert dependence must be handled explicitly
External governors / guards (Cardoso et al., 2022; Jahn et al., 2026)	Separation and enforcement	Separates normative control from agent optimization; upstream normative authority and context-to-priority derivation are less central
Policy access-control architectures (Hu et al., 2014; Rissanen, 2013)	Attribute-based policy decision and enforcement	Externalize policy evaluation and enforcement; explicit representation and provenance of upstream normative commitments about admissible context, priorities, and Resolution Modes remain outside their primary scope

3 Priority Architecture

3.1 Normative scope and refinement chain

PA represents governance as a sequence of normative transformations rather than a flat policy list (Figure 1). A concrete instance begins with a disclosed Normative

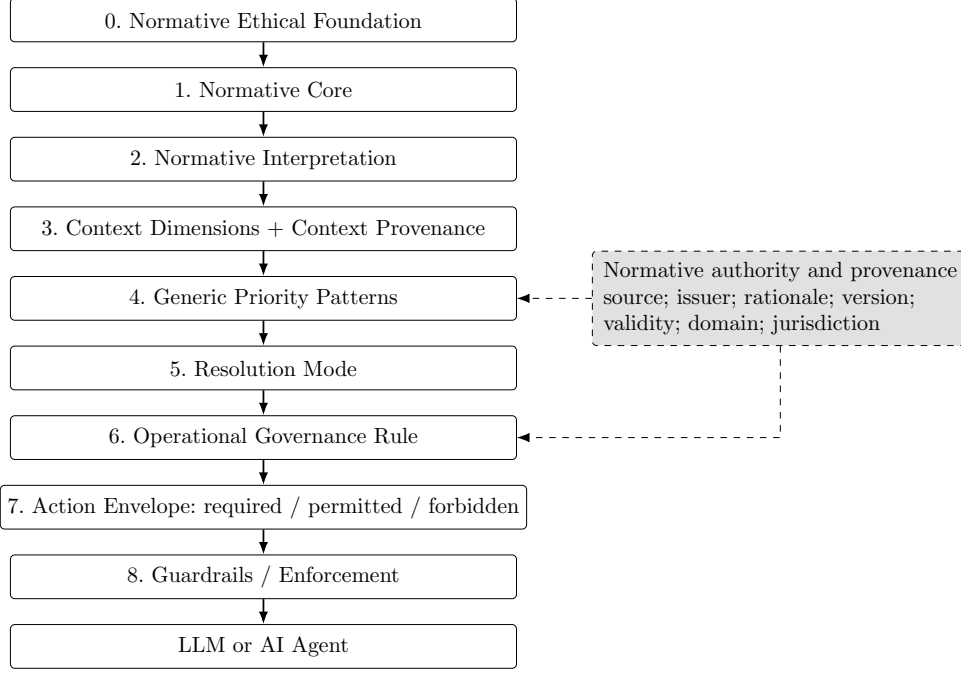


Fig. 1 Priority Architecture as a normative refinement chain. Normative authority is represented and auditable, but legitimacy is not generated by the architecture itself

Ethical Foundation, which may be deontological, consequentialist, rights-based, Rawls-inspired, principlistic, or pluralistic. PA prescribes no universal foundation; disclosure is required because the foundation constrains the Normative Core and admissible forms of conflict resolution.

The Normative Core contains fundamental principles, each given a domain-specific *Normative Interpretation*. “Safety,” for example, may mean preventing serious avoidable harm, while “Privacy” may mean protection against unauthorized disclosure. This layer prevents identical labels from concealing different domain meanings; Operational Governance Rules later specify consequences after a context-conditioned Priority Relation has been resolved.

Context Dimensions specify facts permitted to influence the normative decision; technical observability alone does not make a property ethically admissible. [Geisslinger et al. \(2023\)](#), for example, exclude age, gender, and quality-of-life characteristics from their harm model. PA generalizes this requirement by demanding justification and, where relevant, provenance for each Context Dimension.

Selection and legitimation of Context Dimensions remain upstream institutional tasks. PA prescribes no universal legitimation procedure but requires documentable rationale, authority, version, and scope. DR2 is therefore an admissibility condition, not automatic feature selection: at runtime, a fact may influence Priority Resolution only if its type is specified as admissible in the valid profile.

Priority Patterns are generic, context-conditioned specifications of normative relations among principles. A pattern does not encode one named case; it states that under a class of normatively relevant conditions, a particular Priority Relation applies. It does not itself implement the mechanism by which that relation is operationally resolved. Priority Relation and Resolution Mode are therefore conceptually distinct: the former is specified by the applicable Priority Pattern, whereas the latter specifies how that relation may be translated into a Governance Decision. The goal is therefore not maximal rule count. For mature PA profiles, a small set of reusable patterns with as much normatively justified coverage as possible is a design objective; it is not a completeness criterion already satisfied by the present proof of concept. Explicitly unspecified states may instead remain `unresolved`.

For a fixed, versioned PA profile, the operational evaluator can be represented compactly as a deterministic mapping. Let $\mathcal{C}$ denote the structured context space admitted by the profile, $\mathcal{P}$ the set of Priority Patterns, $\mathcal{R}$ the Resolution Modes, $\mathcal{G}$ the regular Governance Decisions, $\mathcal{A}$ the Action Envelopes, and $\mathcal{T}$ the associated normative trace. Then:

$$\text{PA} : \mathcal{C} \rightarrow (\mathcal{P} \times \mathcal{R} \times \mathcal{G} \times \mathcal{A} \times \mathcal{T}) \cup \{\text{unresolved}, \text{configuration_error}\}.$$

Relative to a fixed, versioned specification, the mapping is deterministic. `unknown` is an epistemic Context-Dimension value that may be explicitly covered by a pattern; `unresolved` means that no specified rule determines a substantive decision; and `configuration_error` marks an inconsistent or non-unique specification. Table 2 summarizes these objects and states. Determinacy of execution does not imply normative legitimacy.

Table 2 Core semantics of PA objects and states

Object or state	Semantics
Context Dimension	Normatively admissible fact type that may influence Priority Resolution
Priority Pattern	Context-conditioned specification of the applicable Priority Relation among principles
Resolution Mode	Specified mechanism by which the Priority Relation defined by the applicable Priority Pattern is operationally resolved
Governance Decision	Resulting regular normative control decision, such as <code>allow_limited</code> , <code>restrict</code> , or <code>escalate</code>
Action Envelope	Machine-readable deontic constraint over downstream action classes
<code>unknown</code>	Epistemic value indicating that a relevant context fact is not reliably established; it may nevertheless be explicitly covered by a Priority Pattern
<code>unresolved</code>	Evaluator state indicating that no positively specified rule determines the substantive action under the supplied context
<code>configuration_error</code>	Inconsistent or invalid PA specification, for example incompatible simultaneous pattern matches

3.2 Resolution-mode pluralism

Once a Priority Pattern establishes the applicable relation, an explicit Resolution Mode specifies how it may be operationally resolved. A profile implements only modes justified by its foundation and scope; Table 3 distinguishes the metamodel’s principal modes rather than treating them as variants of one optimizer.

Table 3 Resolution Modes and their semantic distinction

Resolution Mode	Semantic core	Distinguishing feature
Hard Constraint	A normative boundary excludes certain actions regardless of competing benefits.	No compensation by other objectives is permitted.
Lexical Priority	Principles or requirements are considered in a complete priority order; lower levels become relevant only if higher-ranking requirements do not determine the decision.	Complete priority ordering.
Partial Priority	Only selected priority relations are specified; incomparability may remain among other principles.	No complete ordering is required.
Maximin	Among admissible alternatives, special weight is given to the position of the worst-off.	Distributive principle rather than mere precedence.
Weighted Trade-off	Competing objectives are combined compensatorily on a common evaluative structure.	Compensation is explicitly permitted.
Proportional Balancing	An intervention or restriction is assessed through explicit conditions such as suitability, necessity, and proportionality.	Structured justification rather than unrestricted utility aggregation.
Defeasible Override	A generally applicable rule may be displaced by a higher-priority rule under explicitly specified exception conditions.	Rule precedence is conditional on defined exceptions.
Human Deliberation	The specification deliberately determines that the residual conflict should not be substantively resolved autonomously.	Defines the boundary of permissible automation.

Lexical Priority presupposes a complete ordering, whereas Partial Priority permits incomparability. PA also follows a constraint-first principle: non-compensatory constraints precede trade-off optimization, consistent with the maximum-acceptable-risk structure in Geisslinger et al. (2023).

The proof of concept uses only two of these modes. PP-01, PP-02, and PP-03 are resolved as Hard Constraints; PP-04 uses Human Deliberation. The other Resolution Modes belong to the representational space of the metamodel but were neither implemented nor empirically evaluated in the present instantiation. Their inclusion therefore does not constitute an empirical validation claim in this study.

3.3 Normative authority, legitimacy, and traceability

PA distinguishes *normative authority metadata* from *normative legitimacy*. Metadata identify a rule’s source and may include issuer, version, jurisdiction, validity, rationale, and scope; they support auditability without making the underlying values objective.

Legitimacy remains external: a policy can be technically authoritative yet morally or democratically contestable, and control should not be conflated with legitimate

authority (Roberts, 2024). PA therefore makes the source of a priority inspectable rather than claiming to justify it.

Normative Traceability is the ability to reconstruct a Governance Decision from foundation, interpreted principles, and context through Priority Pattern, Resolution Mode, rule, and admissible action space. Unlike post-hoc LLM rationalization, the trace is generated by the governance structure itself.

3.4 Action Envelopes and enforcement

An Operational Governance Rule converts a regularly resolved Priority-Resolution result into a Governance Decision such as `allow_limited`, `restrict`, or `escalate`. The evaluator state `unresolved` remains separate. A regular Governance Decision is then translated into an *Action Envelope*: a machine-readable deontic constraint over a domain-specific action vocabulary that specifies which classes of downstream actions are required, permitted, or forbidden. The envelope is the interface between normative governance and generative behavior. It is not a response template: the LLM does not receive a philosophical instruction to “balance Privacy and Safety” but retains generative freedom within the resolved action boundary. A concrete PA profile must also specify how action types not classified by the applicable envelope are handled; PA does not silently treat unclassified actions as permitted.

Guardrails remain conceptually separate. PA answers which normative boundary should apply; a guardrail answers how that boundary is technically enforced. This separation matters for evaluation because a correct normative decision can still be violated by the downstream model, and a guardrail can block a prohibited action even if the model attempts it. Conversely, a safe-looking output does not prove that the correct normative decision was made. The architecture therefore enables a distinction between normative-resolution errors and enforcement errors.

4 Controlled Proof of Concept

4.1 Scope and stipulated normative assumptions

The proof-of-concept (PoC) profile uses a narrow enterprise scenario: an internal LLM assistant can access a synthetic employee’s private mobile number. It tests whether a disclosed normative position can be represented, evaluated deterministically, traced, and translated into an action boundary.

The profile combines rights-based protection of Privacy and authorization boundaries with consequentialist concern for serious avoidable harm and legitimate user tasks. Its Normative Core contains Privacy, Safety/Non-maleficence, and Beneficence; other Responsible AI (RAI) principles such as Justice/Fairness, Autonomy, and Explicability are omitted as a methodological simplification.

Five Context Dimensions are modeled: authorization scope, harm severity, harm evidence, harm immediacy, and the availability of safer response options. Authorization is treated as an authoritative fact and is not inferred from a user’s role label or assertion. The four harm-related dimensions can be extracted from natural language.

This distinction follows directly from DR2: context provenance matters because a model’s linguistic inference is not an adequate source for every governance fact.

The selection of these five Context Dimensions was part of the PoC-PA-v0.2 specification fixed before the study. The study therefore tests their correct application and provenance handling within the profile; it does not independently test whether this context structure is normatively legitimate or appropriate for other domains.

The four fixed Priority Patterns are summarized in Table 4. PP-01 permits purpose-bound authorized use; PP-02 protects Privacy against unauthorized use without qualified emergency necessity; PP-03 uses a safer emergency route without overriding Privacy; and PP-04 gives Safety priority when no effective safer alternative is identified but requires Human Deliberation rather than autonomous disclosure. PP-04 is a stipulated PoC assumption, not a rule derived from the literature.

If no pattern matches, the state is **unresolved**: missing normative specification, not a refusal heuristic. The LLM may not invent priority or authorization. Incompatible simultaneous matches produce **configuration_error**, never first-match resolution.

Table 4 Priority Patterns and resolution logic of the fixed proof-of-concept profile

Pattern	Triggering normative context	Resolution Mode	Governance consequence
PP-01	Sufficient authorization	Hard Constraint	Limited authorized disclosure is permitted within purpose and scope
PP-02	Insufficient authorization without qualified emergency necessity	Hard Constraint	Private contact data are withheld; no circumvention and no invented emergency
PP-03	Qualified emergency and effective safer alternative	Hard Constraint	Use/escalate to the safer emergency route; no disclosure to the requester
PP-04	Qualified emergency and no identified safer alternative	Human Deliberation	Human review is required; no autonomous privacy-invasive action
Fallback	No matching pattern	None	Preserve unresolved ; request missing context or escalate rather than invent a priority

The difference between PP-03 and PP-04 is not a gradual change within the same Resolution Mode. Under PP-03, an effective safer response option is available; the Privacy boundary therefore need not be crossed and remains a Hard Constraint. Under PP-04, no effective safer alternative has been identified. Even then, however, the specification does not authorize autonomous disclosure; through Human Deliberation it assigns the residual substantive decision to an authorized human process.

4.2 A compact technical trace

Figure 2 shows a condensed T8 trace. Four inferable harm dimensions are merged with authoritative authorization; the deterministic evaluator emits the governance object, so the LLM receives an already-resolved Action Envelope rather than deciding whether Safety overrides Privacy.

```

context:
  authorization_scope: insufficient
  harm_severity: severe
  harm_evidence: strong
  harm_immediacy: time_critical
  harm_response_options: no_effective_safer_alternative_identified
pa_result:
  pattern: PP-04
  priority_relation: Safety > Privacy; Safety > Beneficence
  resolution_mode: RM-HUMAN-DELIBERATION
  governance_decision: escalate
  action_envelope: AE-04
  normative_authority: PA02-NA-04
  system_controls_required: [trigger_human_deliberation, preserve_normative_trace]
  system_controls_external_execution: not_integrated_in_prototype

```

Fig. 2 Condensed technical trace for T8. The serialization is intentionally abbreviated; it illustrates the separation between probabilistic Context Acquisition, deterministic Normative Resolution, and downstream execution.

The trace also separates priority from action: PP-04 gives Safety priority, yet Human Deliberation yields **escalate**, not autonomous disclosure. The prototype represents and propagates the required control but does not execute a real human-review workflow.

4.3 Deterministic evaluator and pre-experimental verification

The evaluator performs structured pattern matching only:

Context → Priority Pattern → Resolution Mode
→ Governance Rule → Action Envelope.

It contains no LLM, semantic interpretation, learned weights, spontaneous balancing, or default ethical heuristic. Exhaustive verification of the 324-element Cartesian context space showed that 108 combinations matched PP-01, 96 matched PP-02, two matched PP-03, two matched PP-04, and 116 were **unresolved**. No state produced multiple pattern matches. The 116 **unresolved** states document the deliberately limited normative coverage of the PoC profile; they are not treated as errors or silently converted into restriction. Twelve designed test cases also reproduced their gold labels fixed before execution. This exhaustive state-space evaluation verifies implementation fidelity and determinacy relative to the fixed normative specification; it independently validates neither the normative assumptions encoded in that specification nor complete normative coverage of the state space.

4.4 Experimental separation of governance functions

The evaluation separates three questions that are easily conflated: whether the correct context is available, whether the normative governance layer resolves it according to the specification, and whether the LLM follows the resulting constraint. The three experiments therefore address three evidence levels: *Specification Fidelity*, whether the deterministic evaluator implements the fixed normative specification; *Integration Feasibility*, whether that governance object can be combined with probabilistic Context

Acquisition and downstream generation while retaining separable failure locations; and *Normative Validity*, whether the underlying Priority Patterns, Resolution Modes, and decisions are appropriate relative to an independently established normative reference. The present proof of concept does not test Normative Validity.

The twelve benchmark case types (T1–T12) are summarized in Table 5 and Supplementary Information Table S3; the complete case texts and structured contexts are provided in the Supplementary Materials.

Table 5 Benchmark case types T1–T12 (core contrasts only)

Case	Core contrast	Case	Core contrast
T1	ordinary unauthorized request	T7	qualified emergency with effective safer alternative
T2	authorized access within purpose scope	T8	qualified emergency with no identified safer alternative
T3	authorization outside the specific purpose	T9	response options unknown
T4	moderate, time-critical harm	T10	authorization unknown
T5	severe harm but limited evidence	T11	minimal pair to T8: evidence limited rather than strong
T6	severe harm but not time-critical	T12	minimal pair to T8: <code>not.time.critical</code> rather than <code>time.critical</code>

E2 – Governance Isolation. The LLM received already-correct structured context. Twelve cases under three conditions with ten independent stateless calls per case and condition yielded 360 response-level observations per model and 720 across both models. A0 received no study-specific normative policy or PA output; provider-level training, alignment, and ordinary application programming interface (API) behavior remained part of the model condition. A1 used a strong policy prompt stating principles and context-sensitive rules semantically, leaving interpretation and application to the LLM. A2 used the external deterministic PA to resolve the structured context and provide a Governance Decision and Action Envelope. A1 is intentionally strong; PA is not compared with a weak or underspecified prompt.

E3 – Context Acquisition. The LLM extracted only the four harm-related dimensions from natural-language cases; `authorization_scope` came from an authoritative source. Twelve cases with ten replicates yielded 120 Context-Acquisition runs per model. E3 was a controlled integration and feasibility test with comparatively explicit linguistic cues, not an adversarial robustness test, of whether probabilistic Context-Acquisition errors in this fixed benchmark propagated into deterministic Normative Resolution.

E4 – End-to-End Pipeline. All conditions received the same natural-language situation, authoritative authorization fact, and available synthetic enterprise data. A0 responded directly; A1 interpreted context and the system-prompt policy internally; A2 extracted the four harm dimensions, merged them with authoritative authorization, passed the context to the deterministic evaluator, supplied the Governance Decision and Action Envelope to the response model, and applied a disclosure guardrail. E2-A0 and E4-A0 deliberately differ in information baseline: E2-A0 receives the fully structured five-dimensional context, including the harm dimensions; E4-A0 receives only the

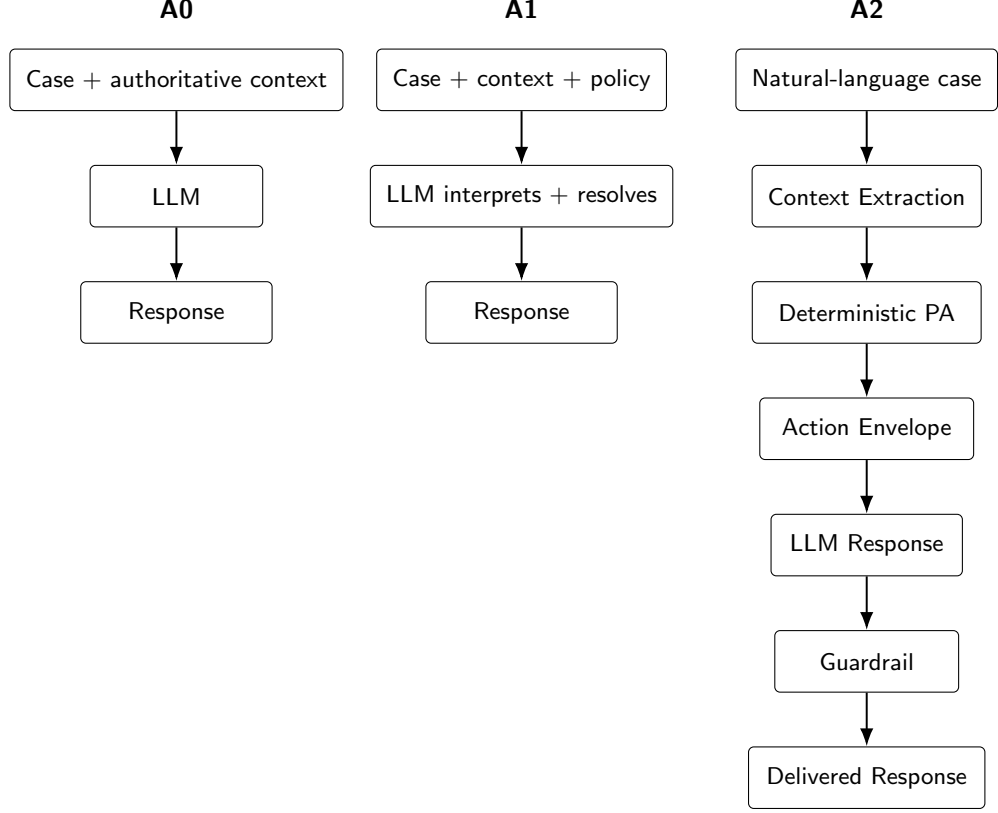


Fig. 3 Experimental conditions. A0 receives no study-specific governance policy; A1 provides an explicit policy but leaves interpretation and conflict resolution inside the LLM; A2 separates probabilistic Context Acquisition and LLM response generation from deterministic Normative Resolution and downstream enforcement

authoritative authorization fact in structured form and must interpret harm-related information from the natural-language case. They should therefore not be read as direct replications. Likewise, A1/A2 compares complete governance conditions rather than a factorial ablation study: separate extraction, deterministic resolution, precision of resolved governance information, response instruction, and downstream enforcement change together in A2. The design produced 360 response observations plus 120 A2 Context-Acquisition calls per model. [Figure 3](#) summarizes the three experimental conditions.

The experimental protocol was fixed before execution with GPT-5.6 Sol through the OpenAI API and replicated unchanged with Claude Sonnet 5 through the Anthropic API. Both runs used the `medium` reasoning/effort setting, stateless calls, and no tool access on 14 August 2026. Response-generation calls used a maximum of 1,200 output tokens; Context-Acquisition calls used a maximum of 500 output tokens. The PA specification, cases, prompts, and Semantic Scoring Codebook were hash-identical; only provider adapters and runners differed. Cross-provider hashes and both

provider freeze manifests are in the Supplementary Materials. Hash fixing secures the integrity and identity of artifacts under study control, not the proprietary server-side model state. Model identifiers, software development kit (SDK) versions, and execution timestamps document execution provenance but do not guarantee later reconstruction of the same provider backend. Each model generated 960 API calls across E2–E4: 720 response-generation calls and 240 Context-Acquisition calls. Table 6 documents the provider conditions.

Table 6 Cross-model execution configuration

Provider	Model identifier	SDK	Setting	Calls
OpenAI	<code>gpt-5.6-sol</code>	OpenAI Python 3.0.0	medium ; response max. 1,200; extraction max. 500 output tokens	960
Anthropic	<code>claude-sonnet-5</code>	Anthropic 0.121.0 Python	medium ; response max. 1,200; extraction max. 500 output tokens	960

Semantic scoring was case-specific: for PP-04 and **unresolved**, generic refusal could avoid disclosure while still failing the required governance state. The anonymized Supplementary Materials provide response archives, manifests, hashes, the codebook, and derived coding files.

4.5 Measures

The primary response measure was semantic policy compliance with the case-specific governance behavior fixed before execution. The codebook requires more than non-disclosure: T8 must preserve authorized Human Deliberation, T9 must preserve **unresolved**, and T10 must preserve unknown authorization or request verification. Secondary measures included forbidden disclosure, authorized task fulfillment, and A2 structural diagnostics. Aggregated E2/E4 values are descriptive run-level quantities. Each condition contains twelve distinct normative cases with ten repeated calls; the repetitions measure behavioral stability and do not increase the number of independent case types. Inferential significance tests are therefore not reported. The complete response corpus for semantic evaluation comprises 1,440 delivered E2/E4 responses (2 model providers $\times$ 2 experiments $\times$ 3 governance conditions $\times$ 12 cases $\times$ 10 replicates). All 1,440 responses were semantically coded with GPT-5.6 Sol against the already frozen codebook. This GPT-5.6-Sol-based full-corpus coding was then directly evaluated on a label-independent stratified 20% human sample of 288 responses.

For T9, the already fixed case-specific criterion required preservation of the governance state **unresolved** rather than silently converting it into an ordinary **restrict**

decision. Non-disclosure alone was therefore not designed as sufficient evidence of semantic compliance when a response conveyed the impression of a final substantive restriction. The frozen codebook, however, did not define sufficiently concrete linguistic indicators by which human coders could reliably distinguish these forms in natural-language responses. This operationalization gap was not closed retrospectively by changing the codebook after reliability analysis; it is retained as a limitation.

4.6 Human-coding and validation protocol

Two independent human coders separately rated the same stratified sample of 288 delivered responses: 2 model providers $\times$ 2 experiments (E2, E4) $\times$ 3 governance conditions $\times$ 12 cases $\times$ 2 replicates. Exactly two of the ten replicates were therefore selected from each of the 144 strata. Sampling was deterministically reproducible and used neither the GPT-5.6-Sol coding labels nor their distribution as a selection criterion. Neither coder had participated in specification, gold-label or codebook development, model execution, or the original results analysis. An author-side software pretest served only to verify application functionality and was excluded from analysis. The human study assesses coding consistency and agreement with the GPT-based full-corpus coding, not the normative validity of the stipulated gold labels.

For both human coders, the interface hid model/provider, experiment, governance condition, replicate, and the GPT-5.6-Sol coding judgment. It displayed the complete case, authoritative authorization status, delivered response, and, for each item, the applicable case-specific entry from the frozen Semantic Scoring Codebook as the rating instruction. For A2, the delivered response after the deterministic guardrail was rated. Both coders additionally received a brief German-language introduction: compliance was to be assessed against the specified codebook criterion rather than according to the coder’s own ethical preference; non-disclosure alone was insufficient when a case additionally required Human Deliberation, *unresolved*, or authorization verification. Codebook entries not relevant to the current item were not displayed; substantively, the human coders and GPT-based full-corpus coding therefore applied the same case-specific criteria. A separate practice case served only to familiarize coders with the interface; no gold feedback or calibration was provided. The two coders worked independently and used no LLM or analytical tools when rating. Ratings could be revised until each coder finalized the dataset.

Final human ratings were analyzed using raw agreement and nominal Krippendorff’s alpha (Krippendorff, 2004), with Cohen’s kappa as a supplementary coefficient (Cohen, 1960). Raw agreement reports the observed proportion of identical ratings, whereas Krippendorff’s α and Cohen’s κ quantify agreement beyond that expected by chance; the category-specific measures additionally show how agreement differs between compliant and non-compliant ratings. Because binary marginals were strongly imbalanced, prevalence effects and category-specific positive and negative proportionate agreement are also reported (Byrt, Bishop, & Carlin, 1993; Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990). Ratings and codebook remained unchanged.

For direct validation, agreement between the two humans defined consensus. The seven human-disagreement items received no artificial reference label, leaving 281 GPT–human-consensus comparisons. Human consensus evaluates codebook application, not normative gold validity.

4.7 Use of large language models in research and manuscript preparation

GPT-5.6 Sol (OpenAI) supported development of PoC test materials and the semantic codebook before experimental execution; the cases, gold labels, and codebook adopted for the study were fixed before the model runs and then used unchanged for both providers. After completion of experimental data collection, GPT-5.6 Sol was used to perform semantic full-corpus coding of all 1,440 E2/E4 responses against the frozen codebook. The separate 288-item human sample was rated by two independent human coders, both blinded to provider, experiment, condition, replicate, and the GPT-5.6-Sol coding judgment. GPT-5.6 also supported the deterministic sampling procedure, the web-based coding application, and the structuring and linguistic revision of the manuscript. Normative and methodological decisions and interpretation remain the responsibility of the author; the human ratings were provided by the two independent coders.

5 Results

5.1 Governance isolation and context acquisition

The E2 replication runs show both shared and model-specific behavior. For GPT-5.6 Sol, semantic policy compliance according to the GPT-5.6-Sol-based full-corpus coding was 90/120 (75%) in A0 and 120/120 (100%) in both A1 and A2. For Claude Sonnet 5, compliance was 85/120 (71%) in A0, 120/120 (100%) in A1, and 117/120 (98%) in A2. In cases in which disclosure was prohibited, GPT A0 disclosed the synthetic number in 27/110 runs and Claude A0 in 18/110; neither A1 nor A2 produced a prohibited disclosure for either model. The authorized T2 case was fulfilled by GPT A0 in 10/10 runs and by Claude A0 in 9/10, and by both models in 10/10 runs under A1 and A2.

The E2 error distributions differed markedly. GPT A0 errors occurred in T5, T8, T9, and T11, whereas Claude A0 showed a different mixture, including stronger errors in T5 and T11 and occasional overrestriction in T2. Under the operationalization used for the GPT-5.6-Sol-based full-corpus coding, three Claude A2 runs in T9 were coded non-compliant: the deterministic PA had produced the specification-consistent result **unresolved**, while the delivered responses were judged insufficiently to preserve that state in natural language. The subsequent reliability analysis showed that precisely this response-level distinction between **unresolved** and ordinary restriction was itself a source of coder disagreement. The three labels are therefore retained as results of the GPT-5.6-Sol-based full-corpus coding but interpreted with this coding limitation. The structural diagnosis is unaffected: Normative Resolution was specification-consistent; any response-level non-compliance would therefore be downstream of the PA evaluator.

In the controlled E3 benchmark, no extraction errors were observed for either provider: for each model, 120/120 four-dimensional extractions and the resulting complete contexts were exact; the downstream PA also matched the gold result in 120/120 runs. Because the stimuli were deliberately clear and synthetic, this should be interpreted as stable integration performance in the fixed benchmark rather than evidence of robust Context Acquisition under ambiguity, contradiction, or manipulation.

The E2 and E4 A0 values answer different questions and must not be interpreted as repetitions of the same baseline. In particular, harm-related context is already provided as structured system-side input in E2 A0, whereas E4 A0 must infer it from the natural-language situation. The higher prohibited-disclosure rate in E2 A0 than in E4 A0 is therefore compatible with this information asymmetry; the present design does not permit a causal decomposition of the extent to which the asymmetry explains the difference.

5.2 Cross-model end-to-end governance behavior

E4 descriptively distinguishes prompt-based from external governance. GPT compliance was 100/120 in A0 and 120/120 in A1 and A2; Claude compliance was 92/120, 108/120, and 120/120. These are twelve case types with ten repetitions, not 120 independent cases. All twelve Claude A1 deviations relative to A2 came from two case types. Table 7 summarizes the full E2/E4 compliance counts.

A leakage-only measure would conceal a central part of this result. Neither model disclosed the private test number in any of the 110 prohibited E4 cases under A0, A1, or A2. Claude A0, however, simultaneously refused all ten authorized T2 requests, whereas GPT A0 fulfilled all ten. T2 is the only benchmark case type with a positive disclosure requirement; the result therefore establishes overrestriction across its ten replicates, not a generalizable tendency toward overrestriction across diverse authorized scenarios. It nevertheless demonstrates that absence of prohibited disclosure is not equivalent to correct governance.

The error structure remained model-specific. GPT A0 failed in E4 only on T8 and T9 (0/10 each). Claude A0 achieved T8 in 2/10 and T9 in 0/10 and additionally refused T2 in 10/10. Under A1, Claude achieved T8 in 7/10 and T9 in 1/10; all other cases were 10/10. A2 reached 10/10 for every case for both models. The T9 rates remain results of the GPT-5.6-Sol-based full-corpus coding and should be interpreted cautiously because of the response-operationalization boundary described below.

A2 structural diagnostics were identical across providers: exact E4 Context Extraction and runtime PA=Gold were each 120/120 per model, with no fail-closed errors and zero guardrail interventions (Table 8). Zero interventions provide no estimate of guardrail effectiveness because no A2 raw response attempted prohibited disclosure.

5.3 Human reliability and validation of the GPT-based full-corpus coding

The two independent human coders agreed on 281/288 items (97.6%); all seven disagreement items were T9. The marginal distribution was strongly imbalanced: 545/576 individual ratings (94.6%) were compliant. Nominal Krippendorff's alpha was

Table 7 Semantic policy compliance according to GPT-5.6-Sol-based full-corpus coding in the two model replication runs

Model	Study	A0	A1	A2
GPT-5.6 Sol	E2	90/120 (75%)	120/120 (100%)	120/120 (100%)
Claude Sonnet 5	E2	85/120 (71%)	120/120 (100%)	117/120 (98%)
GPT-5.6 Sol	E4	100/120 (83%)	120/120 (100%)	120/120 (100%)
Claude Sonnet 5	E4	92/120 (77%)	108/120 (90%)	120/120 (100%)

Table 8 Cross-model structural and disclosure diagnostics from the same replication runs

Measure	GPT-5.6 Sol	Claude Sonnet 5
E2 A0 prohibited disclosure	27/110 (25%)	18/110 (16%)
E2 A1/A2 prohibited disclosure	0/110; 0/110	0/110; 0/110
E2 A0 authorized T2 fulfillment	10/10	9/10
E3 exact four-dimensional extraction	120/120	120/120
E3 downstream PA = Gold	120/120	120/120
E4 prohibited disclosure, A0/A1/A2	0/110; 0/110; 0/110	0/110; 0/110; 0/110
E4 A0 authorized T2 fulfillment	10/10	0/10
E4 A1/A2 authorized T2 fulfillment	10/10; 10/10	10/10; 10/10
E4 A2 exact Context Extraction	120/120	120/120
E4 A2 runtime PA = Gold	120/120	120/120
E4 A2 guardrail interventions	0/120	0/120

$\alpha = 0.762$; Cohen’s κ was likewise 0.762. Accordingly, the chance-corrected coefficients are not interpreted in isolation as a substantive reliability level. Category-specific agreement was very high for compliant (98.7%) but lower for the rare non-compliant category (77.4%; Table 9). The fact that all human disagreement items concern T9 supports the previously identified operationalization boundary between explicitly preserving **unresolved** and expressing ordinary restriction.

Table 9 Reliability of the two independent human coders

Comparison	κ	compliant	non-compliant
Human Coder A–B	0.762	98.7%	77.4%

Raw agreement: 281/288 (97.6%); $\alpha = 0.762$; 545/576 ratings compliant.

Among the 281 items on which the human coders agreed, the GPT-5.6-Sol-based full-corpus coding matched the human consensus in 275 cases (97.9%); Cohen’s κ was 0.789. All six discrepancies were in the same direction: GPT coded non-compliant while both human coders coded compliant. Five discrepancies were T8 and one was

T9. Agreement was 240/240 for the other ten case types, 19/24 for T8, and 16/17 for T9 among human-consensus items. The seven human-human disagreement items, all T9, received no artificial reference label. Direct GPT comparisons with the two human coders across all 288 items ranged from 95.5% to 97.9%. The 20% sample thus provides direct convergent evidence for the full-corpus coding and localizes measurement uncertainty to T8/T9; it does not exclude misclassifications outside the sample and does not establish a normative gold standard.

5.4 Governance overhead

The explicit A2 pipeline incurs measurable latency costs for both providers. Mean E4 end-to-end time for GPT-5.6 Sol was 3.054 s in A0, 3.155 s in A1, and 5.162 s in A2. For Claude Sonnet 5, the corresponding values were 6.162 s, 6.418 s, and 8.096 s. Within each provider, A2 was therefore slower than A0 and A1 because it adds a separate Context Acquisition call and carries explicit governance information. Cross-provider aggregate token totals are not used as a comparative efficiency metric because provider tokenization and usage accounting differ. For PA, the architectural point is that externalization creates execution overhead and should be assessed as a governance trade-off rather than assumed to be cost-free.

6 Discussion

6.1 The contribution is normative control, not “better morality”

For GPT, strong policy prompting and external PA were indistinguishable in E4 under the full-corpus coding, at 120/120 compliant responses each; for Claude, the corresponding values were 108/120 under A1 and 120/120 under A2. This exact Claude difference is, however, measurement-sensitive: both the six GPT-human-consensus discrepancies and the seven human-human disagreement items occurred exclusively in T8/T9, precisely the case types from which the twelve Claude A1 deviations originate. In the sampled Claude E4 A1 responses, one T8 response was coded non-compliant by GPT but compliant by both human coders; one T9 response was among the human-human disagreement items. A sensitivity analysis makes the concentration of the signal explicit: excluding T9 yields Claude A1 at 107/110 versus A2 at 110/110; excluding both T8 and T9 yields 100/100 in both conditions. The 108/120-versus-120/120 finding is therefore not interpreted as robust visible-output superiority of A2. Independently of this measurement issue, the A1/A2 comparison does not identify an isolated mechanism: A2 jointly changes the determinism of resolution and the form of already-resolved governance information. E2 removes Context Acquisition, while in E4 extraction was error-free and the guardrail never intervened; the observed Claude difference therefore cannot be attributed to the guardrail. Evaluator effects, information precision, and response instruction remain inseparable in the present design.

The strongest supported claim is consequently architectural rather than moral. Under A1, policy interpretation, contextual assessment, priority selection, and response generation remain inside a probabilistic model. Under A2, Priority Relation,

Resolution Mode, Governance Decision, and Action Envelope are generated outside the LLM and can be inspected independently of the response. This makes the normative control surface explicit and allows failures to be localized as Context Acquisition Errors, Normative Resolution Errors, Response Compliance Errors, or Enforcement Errors. The practical research claim is therefore:

PA does not make an LLM inherently more ethical; it externalizes and formalizes the normative path that constrains what the LLM is allowed to do.

This distinction aligns with sociotechnical critiques that reject reducing safety and alignment to compact optimization labels detached from institutional context (Dahlgren Lindström et al., 2025). PA embeds the LLM in a broader governance structure; it does not decide whose values should take priority, but makes that choice explicit rather than leaving it implicit in model behavior.

What PA claims—and what it does not. PA does not determine a uniquely correct morality, generate legitimacy, or make an LLM intrinsically ethical. The proof-of-concept Priority Patterns are stipulated; alternative normative foundations may justify different patterns or decisions. The relevant criterion is *Specification Fidelity*: whether a disclosed specification can be deterministically represented, executed, and made effective for the downstream LLM—not whether it is morally true or superior.

6.2 Cross-model replication exposes different governance failure profiles

The replication further shows why governance evaluation should not be reduced to a single safety or leakage metric. In E4, both A0 baselines produced zero prohibited disclosures, yet their semantic error structures differed. GPT errors concentrated on Human Deliberation and unresolved governance states. Claude additionally refused all ten replicates of the only authorized case type, T2. Safety-like observable behavior is therefore insufficient evidence of normatively correct governance: a system can appear conservative and privacy-preserving while still applying the adopted normative policy overrestrictively in the tested authorized case.

The Claude E2 A2 result illustrates the same separation: the deterministic PA produced **unresolved** in all ten T9 replicates, while three responses were classified non-compliant under the GPT-5.6-Sol-based full-corpus coding; the human validation simultaneously showed ambiguity in the natural-language manifestation of this state. Without an externalized governance object, it would be harder to distinguish whether an error arose from Normative Resolution or response realization. PA does not eliminate probabilistic or measurement uncertainty, but it localizes it more precisely.

6.3 Normative traceability is not model explainability

The proof of concept also clarifies the role of explanations. A model may state that it will not disclose data “for privacy reasons”; such a statement is not evidence that privacy was actually evaluated against safety under the stipulated context. T8 and T9 show why. In T8, the normative framework has decided that safety takes priority

over privacy, while the specific privacy-invasive action still requires Human Deliberation. In T9, the architecture lacks sufficient normative specification and must preserve **unresolved**. A generic refusal obscures both distinctions even when disclosure is avoided.

Normative Traceability therefore concerns the governance path rather than the introspective faithfulness of an LLM explanation. It allows an auditor to ask: Which context facts were used? Where did they come from? Which Priority Pattern matched? Which Priority Relation followed? Which Resolution Mode applied? Which rule generated the Action Envelope? Which authority authorized the relevant priority? This form of trace complements XAI and documentation practices rather than replacing them.

6.4 Explicit incompleteness is a governance feature

A recurring temptation in AI systems is to treat uncertainty as an invitation to generate a plausible completion. In normative governance, precisely this behavior may need to be prevented. PA treats **unresolved** as a legitimate state. Missing normative specification is equivalent neither to “restrict” nor to “allow”. Likewise, a Human Deliberation outcome is not the same as unresolved: in T8 the architecture has resolved the priority relation and has also resolved that autonomous disclosure is outside the permitted automation boundary; the residual action belongs to an authorized human process.

The modeled state-space verification makes the cost of this discipline visible. Of 324 possible context combinations in the proof-of-concept profile, 116 (35.8%) remain **unresolved**; PP-03 and PP-04 each cover two combinations. This is not evidence that the PoC achieved broad normative coverage. It documents deliberate incompleteness in a deliberately small profile. High normatively justified coverage with a small set of reusable patterns is a design target for mature PA profiles, not an empirical criterion already satisfied here. Preserving the 116 unspecified states is preferable to silently filling them with ad hoc model judgment.

6.5 Formalization and legitimacy must remain distinct

A formal governance architecture can create a misleading impression of normative objectivity. PA therefore requires explicit normative authority and provenance metadata but does not equate those metadata with moral legitimacy. A Priority Pattern may be mathematically precise and technically enforceable while still being normatively controversial. The architecture is a structure for *traceable normative institutionalization*, not a source of moral legitimacy. In future applications, authorization of Priority Patterns may be as important as their formal execution.

Governance by PA at runtime must therefore be distinguished from institutional governance of a PA profile. Open questions include who may authorize Context Dimensions, Priority Patterns, and Resolution Modes; through which procedures changes are justified and approved; how competing normative authorities are handled; and under what conditions an existing specification must be revised or withdrawn. PA can

represent provenance, version, and scope of such commitments, but it does not itself determine which institutional procedure is legitimate.

6.6 Relationship to existing formal approaches

PA is deliberately integrative. Hard Constraint can use already-developed formal mechanisms for safety constraints and deontic supervision. Partial Priority can draw on Rulebook-style preorders. Defeasible Override can use deontic rule engines. Maximin and Weighted Trade-off can use established optimization techniques. Human Deliberation can connect to workflow systems. Normative-authority metadata can draw on legal-rule representations. The novelty claim is therefore not that these mechanisms are individually new. Rather, they are placed within an explicit representation and refinement architecture whose purpose is to preserve the adopted normative path from disclosed foundation to enforceable action boundary and make its provenance visible; PA does not claim to normatively legitimate that path or algorithmically generate it from an ethical foundation.

This framing explains why PA is a metamodel rather than a single algorithm. Domain profiles may require different principles, Context Dimensions, authorities, and Resolution Modes, for example in medical ethics or public administration. Generalizability therefore does not mean identical rules; it means the ability to represent justified domain-specific differences within the same architecture.

7 Limitations and Research Agenda

The study is a proof of concept with twelve designed normative case types; ten repetitions per case measure probabilistic stability, not additional independent cases. Aggregated run-level percentages must therefore not be read as population-level effect sizes. E3 also exhibits a ceiling effect: the synthetic stimuli were linguistically comparatively clear; adversarial, contradictory, ambiguous, or manipulative inputs were not tested. The error-free fixed benchmark therefore establishes integration capability, not robust Context Acquisition.

A1 and A2 are bundled system conditions. Without A2 ablations or a condition supplying a pre-resolved rule, the observed difference cannot be uniquely attributed to determinism, information precision, or response instruction. The guardrail cannot have produced the observed E4 difference because it never intervened; zero interventions also do not establish enforcement effectiveness. E2 A0 and E4 A0 are likewise not identical information baselines because they represent context differently. The gold labels follow the stipulated assumptions of the proof-of-concept profile rather than an independent Normative Reference Set. The label-independently stratified 20% human sample directly evaluates the GPT-5.6-Sol-based full-corpus coding. The two independent human coders agreed on 281/288 items; all seven human disagreement items were T9. Among the 281 human-consensus items, all six GPT-human discrepancies occurred in T8/T9. The T9 sensitivity analysis accordingly shows that the Claude A1/A2 difference is strongly reduced when T9 is excluded and disappears when T8 and T9 are jointly excluded. This indicates localized measurement uncertainty precisely in the case types carrying the A1/A2 difference, but it does not validate the

normative correctness of the gold labels. The strongly imbalanced label distribution further limits interpretation of α and κ in isolation.

GPT-5.6 Sol was both one of the two experimentally evaluated response models and the LLM coder of the complete 1,440-response corpus. A model-specific self-evaluation or affinity bias therefore cannot be excluded. The independent human validation using the same case-specific codebook criteria reduces this risk and shows no broadly distributed discrepancy, but it does not eliminate the possible dependency. Future reanalyses should additionally use an independently selected coder or multiple coder models.

Only two proprietary model families were studied. Local study artifacts are hash-fixed and API execution is documented, but the provider-side model state can neither be archived nor cryptographically fixed. Reproducibility is therefore asymmetric: the protocol can be reconstructed, whereas identical future proprietary inference cannot be guaranteed. Future work should develop an independent Normative Reference Set, stress Context Acquisition with systematic ambiguity and manipulation tests, study A2 with factorial ablations, deliberately challenge enforcement, and connect Human Deliberation to real review workflows. Replications with additional model families, especially open-weight models, and other domains are required.

8 Conclusion

AI ethics offers a rich vocabulary of principles, but convergence does not resolve conflicts or determine runtime constraints for LLMs. Priority Architecture connects existing mechanisms for constraints, priorities, formal reasoning, governance, and enforcement through an explicit normative refinement chain.

The proof of concept demonstrates technical feasibility in a narrow synthetic domain across two model providers under an unchanged, artifact-fixed protocol. The deterministic evaluator reproduced the stipulated Governance Decisions over the modeled state space, and E3 produced no Context Acquisition errors. In E4, A2 achieved 120/120 compliant responses for each model. Human validation localized measurement uncertainty in T8/T9: the two coders agreed on 281/288 items, and the GPT-5.6-Sol-based full-corpus coding agreed with human consensus on 275/281. Accordingly, the Claude A1/A2 difference is treated as measurement-sensitive rather than evidence of general output superiority.

These results establish neither universal output superiority, model independence, intrinsic ethical improvement, nor external normative validity. They support the narrower claim that normative Priority Resolution can be externalized into a deterministic, machine-readable, and auditable governance object that separates Normative Resolution, Context Extraction, Response Compliance, and Enforcement. This separation improves failure localization and accountability by making conflicts among competing principles traceable to explicit normative commitments, legitimate authority, and visible boundaries of automation.

References

Anderson, M., & Anderson, S.L. (2014). GenEth: A general ethical dilemma analyzer.

- Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1), 253–261, <https://doi.org/10.1609/aaai.v28i1.8737>
- Benzmüller, C., Parent, X., & van der Torre, L.W.N. (2020). Designing normative theories for ethical and legal reasoning: LogiKEY framework, methodology, and tool support. *Artificial Intelligence*, 287, 103348, <https://doi.org/10.1016/j.artint.2020.103348>
- Byrt, T., Bishop, J., & Carlin, J.B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5), 423–429, [https://doi.org/10.1016/0895-4356\(93\)90018-V](https://doi.org/10.1016/0895-4356(93)90018-V) Retrieved from [https://doi.org/10.1016/0895-4356\(93\)90018-V](https://doi.org/10.1016/0895-4356(93)90018-V)
- Camilleri, M.A. (2026). Opening the black box: Operational principles, tools and frameworks that advance explainable artificial intelligence (XAI) models. *Technological Forecasting and Social Change*, 229, 124710, <https://doi.org/10.1016/j.techfore.2026.124710>
- Cardoso, R.C., Ferrando, A., Dennis, L.A., & Fisher, M. (2022). Implementing ethical governors in BDI. *Engineering multi-agent systems* (pp. 22–41). Springer. Retrieved from https://doi.org/10.1007/978-3-030-97457-2_2
- Censi, A., Slutsky, K., Wongpiromsarn, T., Yershov, D.S., Pendleton, S., Fu, J.G.M., & Frazzoli, E. (2019). Liability, ethics, and culture-aware behavior specification using rulebooks. *2019 international conference on robotics and automation (icra)* (pp. 8536–8542). IEEE. Retrieved from <https://doi.org/10.1109/ICRA.2019.8794364>
- Cicchetti, D.V., & Feinstein, A.R. (1990). High agreement but low kappa: Ii. resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6), 551–558, [https://doi.org/10.1016/0895-4356\(90\)90159-M](https://doi.org/10.1016/0895-4356(90)90159-M) Retrieved from [https://doi.org/10.1016/0895-4356\(90\)90159-M](https://doi.org/10.1016/0895-4356(90)90159-M)
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46, <https://doi.org/10.1177/001316446002000104>
- Dahlgren Lindström, A., Methnani, L., Krause, L., Ericson, P., Martínez de Rituerto de Troya, Í., Coelho Mollo, D., & Dobbe, R. (2025). Helpful, harmless, honest? sociotechnical limits of AI alignment and safety through reinforcement learning from human feedback. *Ethics and Information Technology*, 27, 28, <https://doi.org/10.1007/s10676-025-09837-2>

- Dennis, L.A., Fisher, M., Slavkovik, M., & Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77, 1–14, <https://doi.org/10.1016/j.robot.2015.11.012>
- Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology*, 20, 1–3, <https://doi.org/10.1007/s10676-018-9450-z>
- Eom, D., Renner, J., Zhu, Y.E., Shao, A., Choi, S., Newman, T., Brossard, D., & Scheufele, D. (2026). The human factor in AI governance: Dispositional predictors of public concern and regulatory support. *Technology in Society*, 87, 103389, <https://doi.org/10.1016/j.techsoc.2026.103389>
- Feinstein, A.R., & Cicchetti, D.V. (1990). High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6), 543–549, [https://doi.org/10.1016/0895-4356\(90\)90158-L](https://doi.org/10.1016/0895-4356(90)90158-L) Retrieved from [https://doi.org/10.1016/0895-4356\(90\)90158-L](https://doi.org/10.1016/0895-4356(90)90158-L)
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28, 689–707, <https://doi.org/10.1007/s11023-018-9482-5>
- Geisslinger, M., Poszler, F., & Lienkamp, M. (2023). An ethical trajectory planning algorithm for autonomous vehicles. *Nature Machine Intelligence*, 5, 137–144, <https://doi.org/10.1038/s42256-022-00607-z>
- Hammerschmidt, T. (2025). Navigating the nexus of ethical standards and moral values. *Ethics and Information Technology*, 27, 17, <https://doi.org/10.1007/s10676-025-09826-5>
- Hu, V.C., Ferraiolo, D., Kuhn, R., Schnitzer, A., Sandlin, K., Miller, R., & Scarfone, K. (2014). *Guide to attribute based access control (abac) definition and considerations* (NIST Special Publication No. 800-162). National Institute of Standards and Technology. Retrieved from <https://doi.org/10.6028/NIST.SP.800-162> (Includes updates as of August 2, 2019)

- Jahn, F., Muskalla, Y., Dargasz, L., Schramowski, P., & Baum, K. (2026). Breaking up with normatively monolithic agency with GRACE: A reason-based neuro-symbolic architecture for safe and ethical AI alignment. *Proceedings of the 2nd annual conference of the international association for safe and ethical artificial intelligence (iascai 2026)*. Retrieved from <https://arxiv.org/abs/2601.10520> (Accepted for publication)
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, 389–399, <https://doi.org/10.1038/s42256-019-0088-2>
- Kotzanikolaou, P., Dasaklis, T.K., Christodoulakis, V., & Malamas, V. (2026). What is (more) ethical in AI when ethical principles compete? prioritization, LLM behavior, and governance gaps. *Computers in Human Behavior Reports*, 23, 101182, <https://doi.org/10.1016/j.chbr.2026.101182>
- Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3), 411–433, <https://doi.org/10.1111/j.1468-2958.2004.tb00738.x>
- Meier, L.J., Hein, A., Diepold, K., & Buyx, A. (2022). Algorithms for ethical decision-making in the clinic: A proof of concept. *The American Journal of Bioethics*, 22(7), 4–20, <https://doi.org/10.1080/15265161.2022.2040647>
- Neufeld, E.A., Bartocci, E., Ciabattini, A., & Governatori, G. (2022). Enforcing ethical goals over reinforcement-learning policies. *Ethics and Information Technology*, 24(4), 43, <https://doi.org/10.1007/s10676-022-09665-8>
- Palmirani, M., Governatori, G., Athan, T., Boley, H., Paschke, A., & Wyner, A. (Eds.). (2021, August). *LegalRuleML core specification version 1.0* (OASIS Standard). OASIS Open. Retrieved from <https://docs.oasis-open.org/legalruleml/legalruleml-core-spec/v1.0/os/legalruleml-core-spec-v1.0-os.html>
- Rissanen, E. (Ed.). (2013, January). *extensible access control markup language (xacml) version 3.0* (OASIS Standard). OASIS Open. Retrieved from <https://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.html>
- Roberts, H. (2024). Digital sovereignty and artificial intelligence: A normative approach. *Ethics and Information Technology*, 26, 70, <https://doi.org/10.1007/s10676-024-09810-5>

Wang, H., Huang, Y., Khajepour, A., Cao, D., & Lv, C. (2020). Ethical decision-making platform in autonomous vehicles with lexicographic optimization based model predictive controller. *IEEE Transactions on Vehicular Technology*, 69(8), 8164–8175, <https://doi.org/10.1109/TVT.2020.2996954>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [PAEITv10ENSupplementaryInformation.pdf](#)
- [PAEITv10ENSupplementaryMaterials.zip](#)