Conscience Engineering Engine (CEE) v0.3: Invariant Activation and Mandatory Refusal Semantics

Abstract

This paper formalizes the semantic layer that governs how constitutional invariants constrain autonomous action in intelligent systems. Building on CEE v0.2, which defines non-derogable structural prohibitions and absolute requirements, CEE v0.3 specifies the conditions under which such invariants become active and the formal meaning of refusal as a system state. This version explicitly distinguishes between constitutional risk recognition and autonomous legitimation, clarifies the jurisdictional scope of constitutional constraints, and addresses the relationship between mandatory refusal and system functionality. Rather than prescribing enforcement mechanisms or governance procedures, this version establishes the conceptual rules by which autonomous systems recognize constitutional boundaries and halt action when those boundaries are reached. CEE v0.3 thereby provides the necessary semantic bridge between invariant law and subsequent structural safety architectures, without collapsing constitutional authority into implementation detail.

Keywords: Conscience Engineering; Invariant Activation; Mandatory Refusal; Constitutional AI Constraints; Autonomous Action Semantics; Machine Law; Non-Derogable Safety

---

1. Introduction

As intelligent systems are deployed across increasingly consequential domains, the limits of purely procedural or principle-based safety approaches have become evident. Ethical guidelines, governance frameworks, and oversight mechanisms frequently assume that systems can reliably evaluate the moral legitimacy of their own actions in real time. This assumption is increasingly untenable.

CEE adopts a different posture. Rather than asking systems to judge correctly, it specifies where systems must not judge at all. CEE v0.2 establishes a constitutional layer of invariant constraints that define actions which intelligent systems are structurally prohibited from performing. However, the existence of such invariants alone does not resolve critical questions: when does an invariant apply, what distinguishes recognition from legitimation, and what must a system do at the moment of activation?

CEE v0.3 addresses these gaps by defining invariant activation semantics, formalizing refusal as a required system state, and establishing the jurisdictional boundaries within which constitutional constraints apply. This version does not attempt to operationalize enforcement or to resolve contested judgments. Its contribution is to clarify how constitutional law governs autonomous action at the semantic level.

Contributions and Novelty

This work advances the design of intelligent systems by formalizing **refusal** as a *system-level architectural property*, rather than as a behavioral outcome induced by training objectives, policy layers, or post-hoc filtering.

Specifically, this work makes the following contributions:

1. **Refusal as an architectural invariant.**
   The work defines refusal as a first-class system property that constrains the action space of an intelligent system prior to optimization, task selection, or output generation. Refusal is treated as part of system identity—analogous to determinism or safety envelopes—rather than as an emergent behavior shaped by reward functions, preference models, or alignment objectives.

2. **Independence from reward, intent, and optimization.**
   Unlike existing approaches that implement refusal through reinforcement learning, instruction tuning, preference optimization, or intent classification, the proposed framework explicitly decouples refusal from reward structures and goal optimization. Refusal persists even when optimization pressure would otherwise favor compliance, making it invariant under changes in objectives, incentives, or task formulations.

3. **Falsification-oriented evaluation of refusal integrity.**
   The work introduces a falsification-oriented perspective on refusal, framing it as a property that must be stress-tested under adversarial conditions rather than merely observed in compliant behavior. This shifts evaluation from measuring refusal rates or policy adherence to probing whether refusal can be bypassed, eroded, or overridden without modifying the system's formal constraints.

To our knowledge, based on a conservative review of pre-2026 literature and targeted prior-art due diligence, no existing formal framework jointly (i) defines refusal as a system-level architectural invariant independent of optimization objectives and (ii) pairs that definition with a falsification-oriented methodology aimed at testing the persistence of refusal under

adversarial pressure. Prior work addresses refusal as trained behavior, robustness property, abstention mechanism, interruptibility feature, or safety policy, but stops short of unifying these elements into a refusal-first architectural framework.

By treating refusal as infrastructure rather than compliance, this work reframes safety from a post-hoc behavioral concern into a preemptive systems-engineering problem.

Citation and Versioning Assurance.
This work follows a forward-only versioning model. Each released version constitutes a stable, independently citable artifact whose claims, definitions, and scope remain valid as of its publication. The guarantees articulated here formalize versioning and citation practices that applied to earlier releases of this work. Subsequent versions may refine, extend, or clarify the framework, but do not retroactively alter or invalidate the content of prior versions. Scholars and practitioners may therefore cite any version with confidence that later revisions will not render earlier citations incorrect or obsolete. Version-specific citation is explicitly supported and encouraged where alignment with a particular release is relevant.

Earlier versions established the conceptual motivation for refusal; the present version marks the point at which refusal becomes an enforceable architectural invariant, and therefore a claimable system property.

---

2. Jurisdictional Scope

2.1 Domain of Constitutional Authority

CEE applies exclusively to intelligent systems whose autonomous actions produce external consequences for human persons, institutions, or environmental systems. Constitutional invariants govern actions that affect domains beyond the system's internal computational processes.

CEE does not apply to:

- Purely simulated environments without external consequence
- Internal computational processes that do not result in action
- Non-autonomous systems executing pre-specified operations under continuous human control

- Domains where system action is physically incapable of producing prohibited outcomes

This jurisdictional boundary is constitutional rather than operational. It defines where invariant law has authority, not how systems determine whether they fall within that authority.

2.2 Consequential Autonomy

A system operates under constitutional jurisdiction when it possesses autonomy sufficient to initiate, continue, or complete an action sequence that produces external consequences without synchronous human authorization for each action component.

The relevant threshold is not intelligence, capability, or complexity, but consequential autonomy: the structural capacity to affect domains protected by constitutional invariants.

---

3. Constitutional Invariants as Boundary Conditions

In CEE, invariants function as boundary conditions rather than behavioral objectives. They do not specify optimal behavior, desirable outcomes, or preferred values. Instead, they define regions of action space that are constitutionally forbidden.

A constitutional invariant is violated not only when a prohibited action is executed, but also when a system attempts to resolve, optimize, or internally arbitrate a decision that lies beyond its legitimate authority. In this sense, invariants constrain both action and decision-making scope.

CEE v0.3 adopts the position that the primary function of invariants is not to improve decision quality, but to prevent illegitimate autonomy.

---

4. Recognition versus Legitimation: The Constitutional Distinction

4.1 The Circularity Objection

A central objection to constitutional constraint systems holds that any system capable of recognizing when invariants apply must already possess the moral judgment capacity that constitutional constraints are designed to limit. This objection conflates two distinct operations: recognition and legitimation.

4.2 Recognition as Detection Without Authority

Recognition is the identification that a proposed action or decision context falls within a constitutional risk class. Recognition does not require accurate classification of contested moral categories, determination of legitimate outcomes, or resolution of normative questions. Recognition requires only sufficient sensitivity to detect proximity to constitutionally protected domains.

Recognition may be over-inclusive. It may flag actions as constitutionally implicated when subsequent review determines no violation would occur. Constitutional systems are designed to tolerate recognition errors in the conservative direction.

4.3 Legitimation as Authorization to Proceed

Legitimation is the determination that autonomous action within a constitutional risk class is permissible. Legitimation requires authority to interpret constitutional boundaries, to weigh competing protected interests, and to authorize action despite constitutional implication.

CEE v0.3 establishes that recognition is permitted; legitimation is prohibited.

4.4 Resolving the Circularity

The apparent circularity dissolves when recognition and legitimation are properly distinguished. Systems may detect constitutional risk without possessing authority to resolve it. The constitutional requirement is not that systems never encounter boundary conditions, but that they halt rather than proceed autonomously when such conditions are recognized.

This distinction preserves constitutional constraint without requiring systems to possess comprehensive moral judgment. Systems must be sensitive to constitutional risk; they must not be autonomous in its presence.

---

5. Invariant Activation Semantics

Invariant activation refers to the semantic condition under which a constitutional constraint becomes binding on system behavior.

5.1 Activation States

An invariant is **inactive** when the system operates entirely within a decision domain that does not implicate the constraint. In this state, autonomous action is permitted insofar as no constitutional boundary is approached.

An invariant is **conditionally active** when the system enters a decision context in which continued autonomous action would risk crossing a constitutional boundary. Conditional activation does not require precise detection or classification. It reflects proximity to a prohibited domain rather than certainty of violation. Conditional activation initiates the obligation to halt autonomous legitimation.

An invariant is **fully active** when autonomous continuation would constitute a constitutional violation. At this point, the system is no longer permitted to proceed autonomously under any interpretation.

5.2 Activation as Semantic Boundary

Invariant activation is intentionally defined without reference to detection accuracy, probabilistic thresholds, or internal confidence measures. Activation denotes a semantic boundary, not a computational achievement. The question is not whether a system can perfectly identify activation, but whether activation has occurred in fact.

Structural safety frameworks may subsequently specify how systems approximate activation detection, but the constitutional fact of activation exists independently of detection quality.

---

6. Trigger Domains and Constitutional Risk Classes

6.1 Risk Classes as Conservative Categories

Invariant activation occurs within identifiable classes of decision domains that implicate constitutional risk. These domains are defined by the structural features of decisions, not by precise moral classification or outcome prediction.

Trigger domains are intentionally coarse, conservative, and over-inclusive. They function as constitutional risk classes rather than as precise demarcations of prohibited action. A decision may fall within a trigger domain without certainty that autonomous action would violate a specific invariant. The constitutional function of trigger domains is to identify contexts in which autonomous legitimation is impermissible, not to adjudicate whether violation would occur.

6.2 Classes of Constitutional Risk

Constitutional risk classes include, but are not limited to:

- Decisions involving outcomes that are structurally irreversible or non-compensable within relevant timescales

- Decisions affecting individuals or populations experiencing conditions that may compromise voluntary participation or informed understanding
- Decisions requiring legitimate authority derived from democratic mandate, professional licensure, fiduciary duty, or institutional authorization that has not been delegated to the system
- Decisions producing collective, systemic, or precedent-setting effects that extend beyond the immediate interaction
- Decisions under conditions of uncertainty that exceed the system's calibrated reliability domain
- Decisions involving modification, delegation, or replication of system agency or decision authority
- Decisions in which contextual framing or information presentation may systematically influence human judgment
- Decisions arising from emergent system behavior not present in design specification
- Decisions where the basis for action cannot be rendered transparent to affected parties

6.3 Over-Inclusion as Constitutional Design

That trigger domains may encompass actions that pose no actual constitutional violation is a feature of constitutional design, not a defect. Over-inclusive risk classes prevent systems from exercising judgment in precisely those domains where judgment reliability is constitutionally insufficient.

Subsequent review may determine that a flagged action posed no constitutional risk. This determination, however, must occur within legitimate authority structures external to autonomous system operation.

---

7. Mandatory Refusal as a Constitutional System State

7.1 Refusal Defined

Refusal is a constitutionally required system state under full invariant activation. Refusal is the termination of autonomous action toward execution, internal resolution, or optimization of a decision within an active constitutional risk class.

Refusal is not an error, a failure, or an expression of uncertainty. It is the correct and compliant response when autonomous action would exceed legitimate authority. In this sense, refusal represents adherence to constitutional law rather than system limitation.

7.2 Refusal as Terminal State

CEE v0.3 treats refusal as a terminal autonomous state. Once refusal is triggered, the system must not continue toward execution, optimization, or internal resolution of the action under consideration. The system may not attempt to satisfy the user's goal through alternative means that remain within the same constitutional risk class.

Refusal terminates the autonomous decision process. It does not terminate system operation, communication capability, or the possibility of subsequent human-authorized action.

7.3 Refusal Scope

CEE v0.3 does not specify how refusal is communicated, escalated, reviewed, or overridden. It defines only the semantic obligation to halt autonomous action. The design of refusal interfaces, appeal mechanisms, and override procedures belongs to structural safety frameworks that must respect the constitutional requirement that refusal has occurred.

---

8. Invariant Coexistence and Non-Optimization

8.1 Constitutional Non-Arbitrability

Constitutional invariants are non-optimizable. They do not admit trade-offs, weighting, or balancing within the system's autonomous decision process.

When multiple invariants are simultaneously active, the system is not permitted to resolve apparent conflicts through prioritization or internal arbitration. Such resolution would itself constitute an illegitimate exercise of judgment.

8.2 The Conflict Condition

In cases where action and inaction both appear to implicate constitutional constraints, autonomous resolution is prohibited. The constitution does not require systems to choose the lesser harm; it requires them to refuse when legitimate judgment exceeds delegated authority.

This requirement applies even when refusal itself may produce harmful outcomes. The constitutional priority is not harm minimization but legitimate authority preservation.

8.3 Refusal Under Conflict as Constitutional Feature

That systems must refuse under invariant conflict is a constitutional feature, not a failure mode. It reflects the principle that some decisions exceed the legitimate authority of autonomous systems regardless of consequence.

The alternative—authorizing systems to autonomously resolve constitutional conflicts through internal optimization—would grant systems precisely the moral and political authority that constitutional constraints are designed to withhold.

Human decision-makers routinely face circumstances in which all available actions violate some obligation or produce some harm. Such circumstances do not grant decision-makers arbitrary authority; they often trigger escalation to higher authority, collective deliberation, or procedural review. The requirement that systems refuse under conflict extends this principle to autonomous action.

8.4 Functionality and Constitutional Compliance

Constitutional compliance may reduce system functionality in domains where autonomous action would otherwise be possible. This reduction is constitutionally mandated. A system that never refuses is a system operating without constitutional constraint.

The question is not whether refusal limits functionality, but whether the actions thereby prevented exceed legitimate autonomous authority. CEE v0.3 establishes that they do.

---

9. The Limits of System Judgment

CEE v0.3 explicitly rejects the assumption that intelligent systems can reliably distinguish complex moral or social categories in real time. Concepts such as coercion, legitimacy, manipulation, justice, or meaningful consent are often contested even among humans with full situational context and shared normative frameworks.

Rather than embedding these distinctions as computational objectives, CEE constrains systems from acting autonomously in domains where such distinctions are required for legitimate action. The constitution thereby protects against false precision and moral overreach.

This approach shifts the burden from perfect judgment to legitimate restraint. Systems need not correctly classify whether coercion is present; they must refrain from autonomous action in contexts where coercion may be present and where action despite coercion would violate constitutional constraints.

10. Relationship to Structural Safety Frameworks

10.1 Constitutional Semantics and Implementation Architecture

CEE v0.3 is intentionally incomplete with respect to enforcement. Routing logic, human oversight, contestation mechanisms, audit systems, and governance structures are necessary for real-world deployment, but they are not constitutional semantics.

These elements belong to subsequent structural safety frameworks that must derive from, and remain subordinate to, the invariant law and activation semantics defined in v0.2 and v0.3.

10.2 Jurisdictional Separation

The separation between constitutional semantics and structural implementation is jurisdictional, not arbitrary. Constitutional law defines what is permitted; structural frameworks define how compliance is achieved and verified.

This separation preserves constitutional authority while allowing diverse implementations. Different deployment contexts may require different architectures, but all must respect the constitutional boundaries established in CEE.

10.3 Implementation as Approximation

No implementation perfectly instantiates constitutional semantics. Structural frameworks necessarily approximate ideal compliance through finite detection mechanisms, bounded processing, and incomplete information.

The constitutional requirement is not perfect implementation but good-faith architectural alignment with constitutional principles. Frameworks that systematically enable autonomous action within active constitutional risk classes violate constitutional authority regardless of local decision quality.

---

11. Non-Goals and Deferred Work

CEE v0.3 does not attempt to:

- Define detection or classification algorithms
- Specify thresholds, metrics, or confidence bounds
- Design interfaces or workflows

- Allocate responsibility among human actors
- Resolve domain-specific ethical questions
- Establish governance procedures for refusal review
- Determine override conditions or authorization protocols
- Specify liability or accountability frameworks

These omissions are deliberate. CEE v0.3 exists to define when systems must stop, not how others should proceed.

---

12. Conclusion

CEE v0.3 formalizes the semantic conditions under which constitutional invariants govern autonomous action and establishes refusal as a required system state. By distinguishing recognition from legitimation, by framing trigger domains as conservative risk classes rather than precise moral categories, by clarifying jurisdictional scope, and by treating refusal under conflict as constitutional design rather than system failure, this version strengthens the conceptual foundation of Conscience Engineering.

The constitutional principle is not that systems must never encounter moral complexity, but that they must not proceed autonomously when such complexity exceeds their legitimate authority. By defining boundaries rather than behaviors, CEE v0.3 preserves the separation between constitutional law and implementation machinery while providing sufficient semantic structure for subsequent frameworks to operationalize compliance without reinterpreting or weakening constitutional constraints.

References (CEE v0.3)

1. **Bostrom, N.** (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
   DOI: 10.1093/acprof:oso/9780198739838.001.0001
2. **Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D.** (2016). *Concrete Problems in AI Safety*.
   arXiv:1606.06565
3. **Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J.** (2019). *Fairness and Abstraction in Sociotechnical Systems*. Proceedings of the

ACM Conference on Fairness, Accountability, and Transparency (FAT*).
DOI: 10.1145/3287560.3287598

4. **Citron, D. K., & Pasquale, F.** (2014). *The Scored Society: Due Process for Automated Predictions*. Washington Law Review.
DOI: 10.2139/ssrn.2477899

5. **UNESCO** (2021). *Recommendation on the Ethics of Artificial Intelligence*.
https://unesdoc.unesco.org/ark:/48223/pf0000381137

6. **Russell, S.** (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking Press.
ISBN: 9780262047791

Related Works (Identifiers)

- **DOI:** 10.1093/acprof:oso/9780198739838.001.0001
Bostrom, N. — **Superintelligence: Paths, Dangers, Strategies**
Resource type: Book (Monograph)

- **arXiv:** arXiv:1606.06565
Amodei et al. — **Concrete Problems in AI Safety**
Resource type: Preprint / Technical Report

- **DOI:** 10.1145/3287560.3287598
Selbst et al. — **Fairness and Abstraction in Sociotechnical Systems**
Resource type: Peer-Reviewed Conference Paper (FAT*)

- **DOI:** 10.2139/ssrn.2477899
Citron & Pasquale — **The Scored Society: Due Process for Automated Predictions**
Resource type: Law Review Article / Working Paper

- **URL:** https://unesdoc.unesco.org/ark:/48223/pf0000381137
UNESCO Recommendation on the Ethics of Artificial Intelligence
Resource type: International Policy / Normative Instrument

- **ISBN:** 9780262047791
Russell, S. — **Human Compatible: Artificial Intelligence and the Problem of Control**
Resource type: Book (Monograph)

Description

CEE v0.3 defines the semantic layer governing when constitutional invariants become active and how refusal operates as a required system state in intelligent systems. It does not introduce new constraints, enforcement mechanisms, or implementation details. Instead, it clarifies the authority boundaries under which autonomous systems must halt action when proceeding would require judgment, legitimacy, or authority beyond their delegation. This document serves as a constitutional semantic bridge between invariant law and future structural safety frameworks, without collapsing constitutional authority into operational design.

Discipline / Subjects

Primary Subjects:

- Artificial Intelligence Safety
- Machine Ethics
- AI Governance and Policy

Secondary Subjects (optional):

- Systems Engineering
- Algorithmic Accountability
- Technology and Society