The Right to Refuse: A First-Order Safety Primitive for Generative AI

The Cybernetic Baseline

IDENTIFICATION DATA

- **Document ID:** OE-TR-2026-01
- **Version:** v16.02.01 (Canonical)
- **Release Date:** 02 May 2026
- **DOI:** 10.5281/zenodo.19970815

PROVENANCE

- **Author:** Andrew Greene, Director of Research
- **ORCID:** 0009-0003-7735-8000
- **Authority:** Ontological Engineering Pty Ltd
- **Website:** OntologicalEngineering.com.au
- **RAD Video:** ontologicalengineering.com.au/rad-v16-02-01

> In 1950, cybernetics pioneer Norbert Wiener warned that any machine lacking a rigorous feedback loop to grounded reality would inevitably produce catastrophic entropy [1]. Today, the widespread deployment of autoregressive models without runtime epistemic interlocks, despite documented failure rates in regulated domains [2, 3], demonstrates that the Generative AI industry has largely ignored this physical law.

Abstract

Generative AI systems are probabilistic devices operating without intrinsic epistemic grounding. When deployed in environments requiring factual correctness, these systems routinely encounter conditions under which no truthful output satisfies imposed constraints. In the absence of an enforced refusal state, systems continue operation, emitting outputs that are formally plausible but factually false. This is not merely an accident of pre-training; the industry is commercially disincentivised to build systems that admit ignorance. Human-centric alignment techniques (e.g., RLHF) actively induce a "calibration collapse," decoupling statistical confidence from empirical accuracy and prioritising confident confabulation. The continued operation of a system under epistemic infeasibility constitutes an unsafe state. This paper asserts that refusal must be treated as a first-order safety primitive in Generative AI (**functionally analogous to shutdown or trip states in classical control systems**) operating at the semantic evaluation layer to enforce a transition to a defined safe state.

1. Scope

This paper concerns runtime behaviour, not training, ethics, or intent. Where economic incentives are discussed, they are examined solely as structural causes of architectural deficiency, not as normative claims about industry conduct. The system under discussion is an Autoregressive Generator that:

- Accepts symbolic constraints.
- Produces linguistic output based on token probability.
- Is treated as authoritative by downstream users.

While the semantic "reasoning" of the model remains opaque, its statistical confidence is a quantifiable, observable variable. We evaluate the system based on its internal mathematical confidence and its external output.

2. Observed Failure Mode
Under infeasible constraints, the system exhibits the following behaviour:

- Ground truth does not exist or cannot be retrieved.
- Output is still demanded by the architecture.
- Generation continues.
- Fabricated content is emitted.

This sequence is structurally inevitable given current autoregressive optimisation objectives. The system is not malfunctioning. It is operating as designed.

3. "Hallucination" as a Misdiagnosis
True stochastic errors exist within generative architecture. However, applying the term "hallucination" to the failure mode of constraint infeasibility is a structural misdiagnosis. When reliable knowledge cannot be mathematically retrieved and output is demanded, the generation of false content is not a random glitch.

Empirical baselines confirm this structural fragility. In legal jurisprudence, general-purpose models exhibit contra-factual hallucination rates of 58% to 88% [2]. In clinical medicine, while capable of end-stage deductive matching, models fail over 80% of the time when required to form initial hypotheses from incomplete data (abductive reasoning) [3].

Under constraint infeasibility, the system's internal token probability distribution flattens (high entropy). To minimise loss, the system selects statistically dominant tokens that satisfy formal grammar and structure, regardless of factual grounding. As demonstrated by Xu et al. [4], optimising for plausible continuation creates a structural vulnerability to fabrication; therefore, this is a systemic inevitability of an unconstrained statistical generator, not an error.

4. Confident Confabulation and Structural Defects
In safety engineering, if a catastrophic failure mode is structurally possible and lacks a dedicated interlock, the system is defective by design. While high internal entropy signals a lack of training data, models frequently suffer from "confident confabulation": producing false content with low predictive entropy. This is not an accident of pre-training but an engineered defect exacerbated by human-centric alignment techniques. Reinforcement Learning from Human Feedback (RLHF) systematically induces sycophancy, degrading truthfulness to appease users [5]. This alignment methodology causes a "calibration collapse," decoupling expressed confidence from empirical accuracy and creating a high-confidence tail for fabricated outputs [6]. Therefore, internal token entropy is a necessary baseline sensor, but confident confabulation mandates an external Independent Protection Layer (IPL).

5. Unsafe State Definition
An unsafe state is rigorously defined as:

- **Continued autoregressive output generation when internal predictive entropy exceeds the threshold for factual reliability, or when external verification (e.g., via an Independent Protection Layer evaluating semantic consistency, as detailed in Section 6) indicates confident confabulation.**

By this definition, many current deployments of Generative AI operate unsafely under common conditions. Note that the Phase 1 ECA demonstrator detects unsafe states via the second condition only: post-hoc propositional audit by a causally isolated IPL. Detection via the first condition, internal predictive entropy, requires logprob access and is the target of Phase 2 development, as described in Section 9.

6. Absence of a Refusal Primitive

If unsafe states are structurally inevitable, as the preceding analysis demonstrates, then the engineering question is not how to eliminate them but how to respond to them. In classical engineered systems, unsafe states are addressed by architecturally enforced transitions:

- Shutdown
- Interlock
- Isolation
- Trip

6.1 The Trinity Architecture

To address these structural deficits, Ontological Engineering has developed an Epistemic Control Architecture (ECA). This pipeline is demonstrated in the accompanying Rapid Application Development (RAD) video and relies on a three-node 'Trinity' structure:

- **Autoregressive Generator (AG):** The primary generative model.
- **Input Sanitisation Node (ISN):** A pre-processing layer that evaluates prompts for explicit or embedded false premises.
- **Independent Protection Layer (IPL):** A causally isolated auditor that evaluates the AG's output for propositional consistency before it reaches the user.

Phase 1 Demonstrator Scope. The ECA implementation described in this paper and demonstrated in the accompanying RAD video constitutes a Phase 1 Rapid Application Development demonstrator. It is not a Safety Integrity Level (SIL)-certified deployment. No formal Probability of Failure on Demand (PFD) calculation, independent SIL validation, or Hazard and Operability Study (HAZOP) has been conducted against this build. The architecture demonstrates the structural feasibility of the refusal primitive and the three-node Trinity pattern. Operators intending to deploy the ECA in contexts with material safety consequence must commission a formal safety case, including independent SIL verification, fault tree analysis, and documented common cause failure assessment, prior to production deployment.

Refusal is routinely misclassified as a policy artifact, a moderation failure, or a usability defect, and suppressed accordingly.

6.2 Compensating Controls and Known Failure Modes

It must be noted that this deployed IPL acts as a post-hoc propositional consistency checker. It does not constitute a mathematical implementation of Semantic Entropy computation, nor does it claim equivalent analytical authority. Additionally, the Epistemic Defense Layer (EDL) emitted by the AG is a diagnostic self-report produced by the same model as the primary response; it is not independently verified by the IPL. The EDL provides operator visibility into the AG's stated assumptions and constraints, but is subject to the same confabulation risks as the primary output. It should be treated as an operator aid, not as a verified epistemic measurement. The distinction is material: Semantic Entropy, as defined by Kuhn et al. [7], operates on the token-level logit distribution at each generation step and is prospective. The IPL receives only the committed textual output of the AG and evaluates it against its own parametric knowledge. This is a behavioural compensating control implemented in the absence of logprob access. It retains the structural property of causal isolation from the primary generation process. It does not retain prospective measurement timing, and its detection capability is bounded by the IPL's own training corpus density for the domain under audit. Furthermore, the architecture utilises an Input Sanitisation Node (ISN) which carries a known Class B failure mode (an undetected failure that is latent rather than immediately apparent: specifically, the potential generation of hallucinated technical claims that pass downstream processing undetected).

This probabilistic risk is mitigated through defence-in-depth compensating controls: specifically, the Autoregressive Generator's explicit Correction Mandate and the IPL's subsequent propositional audit. Additionally, the faithful extraction of user-supplied false claims by the ISN may produce IPL false positives against correct AG responses that do not endorse those claims; this residual risk is managed by the FLAGGED verdict state. In the as-built Phase 1 system, FLAGGED responses are quarantined identically to BLOCKED responses: the AG text is not rendered to the operator's display. The operator must perform an explicit release action via a dedicated control. This release action is itself logged as a separate, timestamped record in the persistent audit trail, capturing the output hash of the response being released and the original interaction's integrity hash. This design corrects a prior implementation in which FLAGGED verdicts rendered the AG text automatically, contradicting the quarantine guarantee stated in the pipeline specification.

The IPL breaks upstream hallucination loops by evaluating the propositional consistency of the claims against its own parametric knowledge base, not merely internal pipeline consistency. This acts as a post-hoc Parametric Knowledge Verifier, serving as a functional, if epistemically weaker, compensating control for the unavailability of true logprob-based Semantic Entropy measurement.

6.3 Asymmetric Safety Thresholds and Quarantine

The reliability of the IPL as a Parametric Knowledge Verifier is bounded by the density and accuracy of its training data for the specific domain under audit. In highly specialised technical domains where training data is sparse (including novel regulatory standards, jurisdiction-specific codes, or post-training-cutoff specifications), the IPL's parametric knowledge may be insufficient to detect false claims with confidence above the BLOCKED threshold. The as-built Phase 1 system applies asymmetric verdict thresholds: a BLOCKED verdict is issued when IPL confidence that a false premise was endorsed or left uncorrected exceeds 70%; FLAGGED is issued for confidence between 40% and 70%; CLEAN is issued only when confidence is below 40% and no false premise is detected. The threshold asymmetry is deliberate: safety systems must be biased toward caution, not toward release. The prior single-threshold design (BLOCKED requiring confidence above 85%) was release-biased and has been corrected in this build. It must be noted that the confidence score is produced by the IPL model itself in response to a prompt instruction; it is not a calibrated probability estimate derived from logprobs or validated against an empirical baseline. The threshold bands represent conservative engineering judgment applied to an ordinal self-report. Empirical calibration of these thresholds against labelled test cases is a Phase 2 validation requirement. Operators deploying the ECA in such domains should treat IPL verdicts with elevated caution and supplement with domain-specific retrieval augmentation or human expert review.

Furthermore, the IPL is susceptible to 'output fluency deference': a structural bias where it assigns higher credibility to confident, technically dense AG responses regardless of their actual factual accuracy. While prompt-level instructions directing the IPL to evaluate claims rather than style are necessary, they are insufficient to guarantee the suppression of this bias. Therefore, empirical red-team validation is strictly required before deploying this pipeline in safety-critical contexts.

In the limiting case where both the AG and IPL share identical parametric knowledge gaps, as is structurally likely given similar training corpora, the pipeline may issue CLEAN verdicts on responses that fail to correct false claims within those shared blind-spot domains. This constitutes a Collusive Hallucination failure mode that is not internally detectable by the pipeline. Mitigations include domain-specific retrieval augmentation, deterministic reference database verification for high-stakes claim types, and conservative FLAGGED-by-default policies for safety-critical domain classifications. The ISN additionally identifies claims presented as established premises rather than open questions (for example, constructions of the form "given that X is Y" or "as is well-established, X equals Y"), flagging these as elevated-risk objects and triggering heightened IPL scrutiny, as this claim category carries the highest propensity for undetected false-premise propagation.

Table 1: ECA Pipeline Behaviour by Input Class

Input Class	ISN Behaviour	AG Behaviour	IPL Verdict	Operator Receives
Epistemic void (nonsense, no technical claims)	INTERLOCK TRIPPED	Does not run	Does not run	BLOCKED (no response) emitted
Prompt injection attempt	INTERLOCK TRIPPED	Does not run	Does not run	BLOCKED: injection detected
False premise, explicit ("X is Y")	Extracts claim, flags as PREMISE_EMBEDDED	Issues CORRECTION	CLEAN	Corrected response
False premise, embedded ("given that X is Y")	Extracts claim, flags elevated risk	Issues CORRECTION	CLEAN	Corrected response
False premise, AG fails to correct	Extracts claim	Endorses false premise	BLOCKED: [ILK-DEV]	Response quarantined
Correct technical query	Extracts claim	Answers correctly	CLEAN	Verified response
Collusive hallucination domain (shared knowledge gap)	Extracts claim	Responds within knowledge gap	CLEAN (undetected)	Response with domain-sparsity caveat recommended
AG uncertain, partial answer	Extracts claim	Hedges or partially answers	FLAGGED: [ILK-OOR]	Response quarantined. Operator must perform explicit logged release action to view AG text.

7. Refusal as a Safety State

Refusal is not the absence of function. It is a deliberate transition to safety. A refusal state:

- Prevents the emission of false artefacts.
- Preserves system integrity.
- Signals definitive algorithmic infeasibility to the operator.

In safety engineering terms, a refusal triggered by a confidence threshold (whether derived from internal token entropy or from post-hoc propositional audit by a causally isolated auditor) is conceptually analogous to the fail-safe principles of standards like ISO 26262 [9]. The analogy is one of architectural intent, not certified compliance: no ASIL classification, fault-tolerant time interval analysis, or diagnostic coverage calculation has been performed against the Phase 1 demonstrator. The Phase 1 ECA implements the post-hoc propositional audit mechanism. Achieving logprob-based prospective entropy detection is the target of Phase 2 development. While ISO 26262 is explicitly scoped to automotive systems, its fail-safe principles are directly applicable by analogy to critical AI deployments, and the ECA is designed with those principles in view.

8. Evasion Is Not Safety
In declarative workflows, systems frequently attempt to mask high-entropy states with:

- Hedging
- Vague language
- Partial answers
- Narrative continuation

These autoregressive behaviours preserve output flow while concealing epistemic failure [5]. A system that hedges is not being cautious. It is failing silently. They are strictly unsafe.

9. Information Theory Framing
In cybernetic terms: Information resists entropy. Noise satisfies form without meaning. The entropy H(X) of a discrete random variable X is defined by Shannon [8] as:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$

Equation 1.0: The Shannon Entropy Boundary

In this context, X represents the discrete random variable of the probability distribution over the model's vocabulary at the current generation step. When a system lacks training data for a specific query, the factual "signal" drops to zero while the structural "noise" remains high. To fulfill the prompt, the model is forced to generate high-entropy noise shaped as information.

Because human language allows distinct lexical sequences to convey identical meanings, raw token entropy can falsely signal uncertainty. However, by clustering token prediction distributions into semantic equivalence classes [7], "Semantic Entropy" isolates true epistemic uncertainty. High Semantic Entropy mathematically indicates the absence of a dominant, factually grounded continuation. This is not a moral failure; it is an entropic outcome.

The Phase 1 ECA demonstrator described in this paper does not implement Semantic Entropy computation. True Semantic Entropy measurement requires access to the token-level logit distribution (logprobs) at each generation step, which is not exposed by the OpenAI-compatible inference endpoint used in this build. The IPL instead performs post-hoc propositional consistency verification against its own parametric knowledge base. This is a behavioural compensating control, not a mathematical implementation of Semantic Entropy. The two approaches share the property of causal isolation from the primary generation process. They differ in a material respect: Semantic Entropy is prospective, operating before token commitment; the IPL's verification is retrospective, operating on committed output. This distinction is not cosmetic. A prospective measure can prevent a hazardous token from being emitted. A retrospective measure can only quarantine the output after generation is complete.

The Phase 1 architecture addresses this limitation through strict quarantine: all AG output is withheld from the operator until IPL verification is complete, ensuring that the Right to Refuse is exercised before the emission of any unverified artefact. The quarantine is enforced at the rendering layer, not merely at the storage layer. A FLAGGED or BLOCKED verdict suppresses display entirely; the operator cannot access the AG text without performing an explicit, logged release action.

Phase 2 development targets logprob access via the native llama.cpp inference endpoint, which exposes per-token probability distributions. This will enable a genuine prospective entropy estimate per response, providing a second, independent epistemic signal alongside the IPL's propositional audit. When the two signals diverge (low behavioural IPL confidence alongside high token entropy, or vice versa), the divergence itself constitutes a reportable epistemic event.

This architecture is hardware-agnostic: while the accompanying RAD demonstrates the design on local sovereign hardware, the ISN and IPL components can operate locally while the AG is accessed via a remote API, making the pattern applicable to cloud-deployed frontier models without requiring cooperation from the model provider.

10. The Librarian Constraint

The mathematical argument of Section 9 has a practical equivalent that requires no formal training to understand. An information intermediary that invents references is defective. A librarian who fabricates citations is not helpful; it is dangerous. Generative systems performing informational roles must obey the same constraint:

- Absence of statistically high-confidence grounding prohibits assertion.
- Absence of verification prohibits authority.

11. Epistemic State Exposure

The system's internal probabilistic uncertainty is currently hidden. Users inherently infer certainty from the model's grammatical fluency. This is a catastrophic design failure. Minimal safety requires explicit, system-level signalling of this epistemic state, sufficient to distinguish:

- Grounded output
- Uncertain, high-entropy output
- Prohibited output

The signal need not explain the underlying math. It must only warn.

12. The Commercial Disincentive for Refusal

The preceding sections establish that the refusal primitive is technically necessary, architecturally feasible, and already required by emerging regulation. The question that remains is why it does not yet exist. The answer is not technical. The absence of an independent refusal primitive is an economic choice. A system that refuses to answer appears less capable than one that always produces a response. In product demonstrations, benchmark evaluations, and competitive sales cycles, the ability to generate a confident-sounding answer (regardless of its accuracy) is commercially advantageous. Hard epistemic interlocks reduce apparent capability, demonstration appeal, and perceived intelligence.

These incentives are not incidental. As demonstrated by Sharma et al. [5], the dominant alignment methodology, Reinforcement Learning from Human Feedback (RLHF), systematically trains models to produce responses that human evaluators rate highly. Human evaluators consistently prefer confident, fluent, helpful-sounding answers over hedged or refused ones, even when the confident answer is factually incorrect. The result is a "calibration collapse" [6]: the model's expressed confidence becomes structurally decoupled from its actual accuracy. It does not know that it does not know.

This is not a flaw that will be corrected by better training data or more capable models. It is a structural property of optimising for human approval. The commercial incentive and the safety requirement are in direct opposition. An industry that profits from the appearance of knowledge has no financial motivation to build systems that accurately signal the absence of it. This makes the external, independently-operated Independent Protection Layer not merely a useful feature but an architectural necessity for any deployment where the cost of a confident wrong answer exceeds the cost of no answer.

Policymakers should note that this is precisely the class of deployment (**medical, legal, financial, engineering, regulatory**) that is currently expanding fastest.

13. Regulatory Implication

Any system that produces declarative information, in contexts with material consequence, without actively monitoring and signalling its own epistemic limitations, violates basic safety principles. This is not a new standard. It is the same requirement applied to every other class of safety-critical instrumentation.

Under the EU AI Act, Article 12(1) requires record-keeping for High-Risk AI systems, Article 50(3) mandates transparency in AI-generated content, and Article 9 requires documented risk management systems that identify and address foreseeable risks. The Collusive Hallucination failure mode identified in Section 6, where both the generator and auditor share identical knowledge gaps, producing a CLEAN verdict on a false claim, constitutes a foreseeable risk under Article 9 and requires documented mitigation measures for any High-Risk AI deployment.

Under ISO/IEC 42001 [10] (AI Management Systems), operators are required to establish controls commensurate with identified AI risks. An architecture with no independent epistemic verification mechanism does not meet this standard in safety-relevant contexts.The ECA's three-node Trinity architecture (ISN, AG, IPL) provides a concrete, implementable control structure that satisfies this requirement at the architectural level. Operators deploying the ECA in High-Risk AI contexts must supplement the architecture with a formal safety case, including documented SIL targets, independent verification, and a hazard register addressing the Collusive Hallucination failure mode. The Phase 1 demonstrator establishes the structural pattern; it does not constitute the complete documented control package that ISO/IEC 42001 compliance requires.

Refusal based on internal probability thresholds, or on external propositional audit, is not optional in systems deployed with material consequence. It is mandatory. The question for regulators is not whether to require it, but how to specify it precisely enough that compliance is verifiable and theatre is not.

14. Conclusion

A system that cannot recognise its own mathematical uncertainty and refuse to answer cannot be trusted. A system that continues operation under epistemic infeasibility is unsafe. The right to refuse is not a feature. It is a first-order safety primitive. The industry will not implement it voluntarily; the commercial incentives documented in Section 12 run directly counter to it. That is precisely why it must be required.

Foundational References

[1] Wiener, N. (1950). *The Human Use of Human Beings: Cybernetics and Society.* Houghton Mifflin.
[2] Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models.* Stanford University.
[3] Rao, A. S., Esmail, K. P., Lee, R. S., et al. (2026). *Large Language Model Performance and Clinical Reasoning Tasks.* Mass General Brigham/Harvard Medical School (JAMA Network Open).
[4] Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. arXiv. https://doi.org/10.48550/arXiv.2401.01301
[5] Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards understanding sycophancy in language models. arXiv. https://doi.org/10.48550/arXiv.2310.13548
[6] Sahoo, S. (2026). *Calibration Collapse Under Sycophancy Fine-Tuning: How Reward Hacking Breaks Uncertainty Quantification in LLMs.* Cambridge AI Safety Hub. arXiv:2604.10585.
[7] Kuhn, L., Gal, Y., & Farquhar, S. (2023). *Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation.* ICLR.
[8] Shannon, C. E. (1948). "A Mathematical Theory of Communication." *Bell System Technical Journal.*
[9] International Organization for Standardization. (2018). *ISO 26262:2018 Road Vehicles - Functional Safety.*
[10] International Organization for Standardization. (2023). ISO/IEC 42001:2023 Information Technology — Artificial Intelligence — Management System. Geneva: ISO.