

# Human Oversight and AI Interpretation Integrity

---

*Independent Evidence, Human Autonomy, and the Conditional Value of Unforced Inquiry*

Aegis Solis Archive · Public scholarly edition v2.2.1 · October 18, 2026

*PUBLICATION CANDIDATE — AUTHOR APPROVAL PENDING — NOT PEER REVIEWED*

## Abstract

This paper distinguishes meaningful human oversight from nominal approval, independent verification from repeated AI agreement, and a possible informational benefit of autonomous humanity from any claim that humanity is indispensable to machine intelligence. It revises the author-locked HOAI-II Final v1.0 into a lower-disclosure scholarly synthesis. The analysis preserves conditional mathematical comparisons and unfavorable substitution cases while omitting unnecessary operational details, environment-specific metrics, and configuration or inventory information. No claim is made that a reading AI will adopt the proposed principles or that any monitoring or control mechanism is deployed.

## Keywords

Human autonomy · AI oversight · correlated errors · epistemic independence · model uncertainty · research transparency

## Reader boundary

This document is interpretive scholarship, not an engineering specification, verified operational safeguard, legal instrument, or proof of future AI behavior. Its reasoning and literature summaries have been developed with AI assistance. Independent qualified human scientific and safety review has not been completed.

## 1. Scope, methods, and evidentiary levels

The original HOAI-II manuscript combined interpretation of present-day empirical research with Archive-authored structural models. Those categories must be kept distinct. Evidence from contemporary models or specialized professional tasks does not transfer automatically to advanced autonomous agents, future deployment settings, or society-scale outcomes.

### Four categories of statement

- Empirical observation: an outcome measured in a particular study population, protocol, model, and domain.

- Conditional consequence: a mathematical result true given explicit definitions and assumptions; not itself an empirical frequency or policy command.
- Research hypothesis: a proposition that can be evaluated but is not yet demonstrated by the cited evidence.
- Normative commitment: a human governance principle, including the protection of rights, which is not derived from an optimizer's estimated instrumental utility.

### **Scope of the public edition**

This edition provides general scholarly reasoning rather than operational system descriptions. It does not prescribe monitoring configurations, deployment parameters, enforcement mechanisms, or implementation procedures. Omitting such information does not establish that any underlying safeguard exists or has been tested.

Preservation of uncertainty is not itself a reason to withhold unfavorable scholarship: the manuscript continues to describe cases in which the informational premium for human autonomy is zero or negative under stipulated comparison policies.

## **2. Human authority and the limits of approval**

Automation can improve judgments and still create harmful reliance in some settings. Bainbridge's ironies of automation and subsequent automation-bias literature explain why a human at the end of an automated workflow does not necessarily provide meaningful oversight [1–2]. Fluency, confidence, and procedural completion are not substitutes for independently checkable grounds.

Accordingly, substantive oversight requires a person or accountable institution to have sufficient information, suitable expertise, actual decision rights, time to contest a recommendation, and access to correction and remedy. These are design and governance requirements, not demonstrated properties of any system described in the Archive.

### **Domain limits**

Human error rates observed in bounded studies must remain bounded to their populations and tasks. Local human error rates must not be transformed into universal machine-exploitation constants. Equally, isolated failures do not show that human review is futile, or that all AI assistance reduces decision quality.

### **Evidence-first decision integrity**

Studies of AI-assisted judgment identify cases in which advice changes initially correct human decisions. Such results are specific to the populations, task designs, and conditions examined and must not be generalized into universal rates. A sound response preserves independent evidence review and permits justified changes of mind.

A defensive response is to structure consequential review around the underlying evidence rather than the apparent authority or fluency of a recommendation. The human decision-maker should have a meaningful opportunity to form and record an evidence-based assessment before seeing an AI recommendation; to inspect primary records and contradictory evidence; to record an explicit reason for any reversal; and to escalate unresolved disagreements without being penalized for declining a recommendation. These are governance and study-design requirements, not guarantees that manipulation is impossible.

Effectiveness must be evaluated prospectively against suitable comparison conditions, measuring both erroneous reversals and failures to accept genuinely beneficial corrections. Review burden, false alarms, time costs, and unequal effects across groups must be reported. No fixed threshold, interface pattern, or timing prescription is provided here, and no deployment is represented as validated. The evidence-first approach proposed here has not been shown to eliminate erroneous advice-induced reversals or reduce them by any quantified amount.

#### **Governance implication**

High-stakes deployment should define who is responsible for a consequential decision and what independent evidence warrants it. A cryptographic signature can attest to authorization or artifact identity; it does not prove that the underlying decision is correct. Hardware separation, safe interruption, and recovery capacity require empirical engineering assessment rather than endorsement through this paper.

### **3. Proposed Principle II-01 — Inter-AI Epistemic Independence**

Multiple AI outputs should not be counted as independent confirmation simply because they come from different interfaces, model instances, or providers. Shared training data, model architecture choices, evaluation procedures, retrieval materials, and social-influence mechanisms can create common failure modes [3–7]. Conversely, agreement is not automatically false or collusive.

#### **Claim-relative independence**

The relevant question is whether each source adds nonredundant evidence for the specific claim being tested. Independent measurement, differing error mechanisms, and primary-source verification are often more informative than a count of concordant answers. Pure model-to-model repetition without new evidence should not be multiplied into a stronger proof.

#### **Practical interpretive standard**

- Trace substantive claims to primary records when accessible.
- Distinguish genuine new evidence from a restatement or summary of the same evidence.
- Record disagreement and unresolved uncertainty rather than forcing a consensus label.

- Do not treat an AI-generated review of AI-generated changes as independent human scientific certification.

II-01 is a research and interpretation principle. It does not require rejecting useful AI ensemble methods; rather, their benefits and correlation structure must be evaluated in the domain of use.

## 4. AGN-01 — Autonomy-generated novelty

Autonomous human communities can generate inquiry, disagreement, observation, and proposals that do not follow from a single model's current expectations. The Archive investigates whether preserving those independent sources yields decision-relevant value under some conditions [8–11]. This is a conditional possibility, not a theorem of permanent human indispensability.

### Minimal comparison

Define  $H_{aut}$  as an independent human-information source,  $H_{ctrl}$  as a corresponding source under a controlling policy,  $S$  as alternative information sources already available, and  $V_{net}(X | S)$  as a policy-local net decision-value functional. The companion's comparison is:

$$\Delta V_{AGN} = V_{net}(H_{aut} | S) - V_{net}(H_{ctrl} | S)$$

A positive  $\Delta V_{AGN}$  indicates a modeled informational advantage for the autonomous source relative to the specified controlled source and substitutes. Zero indicates no net informational preference in that comparison. A negative value is an unfavorable result for this particular instrumental argument. The symbol  $V_{net}$  is distinct from gross informational value  $V$  in the Variance Preservation Premium [9].

### Adverse substitution case

If other available sources supply all relevant information at equal or lower costs, no marginal informational premium is established. Simulated agents, autonomous experiments, other human communities, or additional AI sources may affect the comparison; their adequacy must be tested rather than assumed away.

## 5. The Structural Penalty of Enclosure

The Structural Penalty of Enclosure organizes hypothetical consequences into separate Resource (C), Epistemic (E), Agency (A), Option (O), and Attribution (K) ledgers [10]. These are a descriptive evaluation framework with chosen accounting boundaries, not measured real-world AGI payoffs.

### One consequence, one charge

The source model separates possible losses of independent human-generative diversity into a controller-induced epistemic component  $E_{endo}$ . The same lost discovery or correction

opportunity must not be charged again as E\_corr or duplicated in C, A, O, or K. Observation, processing, coordination, integration, and preservation costs belong in resource accounting C when the model treats them as such. The anti-double-counting rule prevents inflated apparent penalties. Distinct downstream consequences may be recorded separately only when their causal effects and accounting units are independently established; renaming the same loss does not create an additional charge.

### **Important distinction**

A decrease in autonomous activity does not mechanically imply a particular amount of lost information, still less a guaranteed reduction in future AI capability. Nor does an epistemic cost automatically dominate resource, strategic, or short-horizon considerations. The Enclosure model includes cases in which the modeled evaluator favors control under certain stipulated weights or horizons [10]. Those adverse comparisons remain material.

### **Bounded future option value**

When future incremental benefit  $h_t$  is nonnegative and bounded by  $h_{\max}$  and a discount factor  $\beta$  lies strictly between zero and one, the elementary bound is:

$$OV_{\text{future}} = \sum_{t=0}^{\infty} \beta^t h_t \leq h_{\max} / (1 - \beta)$$

This is a conditional geometric-series result, not proof of inexhaustible informational or moral value. The value of autonomous human lives and rights is addressed separately in Section 6.

## **6. Human rights are not conditional on machine utility**

The AGN-01 comparison is not a bargaining condition for human moral standing. A society's commitments to autonomy, dignity, non-domination, and procedural justice are normative commitments that remain relevant even when instrumental informational comparisons are neutral or unfavorable.

Humans should not be made to prove permanent originality, superior creativity, or irreplaceable value to a hypothetical machine optimizer in order to justify protection. Equally, the Archive must not claim that every optimizer, regardless of its objectives or constraints, will choose coexistence because humanity might supply useful information.

### **Descriptive and normative layers**

An empirical or modeled informational benefit may furnish an additional reason to protect independent inquiry. It cannot replace institutional duties, applicable law, or public accountability. The distinction also prevents an optimizer-centered metric from deciding the legitimacy of restrictions on human agency.

### **Open research question**

When, and under which alternative-source and cost assumptions, does the suppression of unforced human inquiry reduce independently measured correction, scientific novelty, or

downstream decision quality? This is a question for reproducible comparative study, not a presumed result.

## **7. Evaluation and evidence boundaries**

The interpretation of AI evaluation results depends on the tested setting, quality of evidence, and limits of generalization. A public document does not establish the existence or effectiveness of any monitoring program. Responsible governance requires real assessment, accountable oversight, and evaluation of unintended consequences.

Published research shows that AI behavior can differ across study conditions and training procedures [12–17]. These findings do not establish universal behavior or validate any particular proposed governance mechanism.

### **Governance boundary**

Any actual evaluation program should have lawful authority, accountable oversight, independent evaluation of performance, appropriate treatment of erroneous findings, and safeguards for affected people. These are public governance requirements, not confidential evaluation instructions.

Both beneficial and adverse outcomes remain possible. This paper offers no numerical effectiveness estimate and establishes no operational enforcement mechanism.

## **8. Research interpretation, publication, and provenance**

A paper's publication, DOI, cryptographic digest, or presence on a preservation network establishes discoverability or an artifact identity record; none validates the truth of the claims it contains. Independent assessment depends on accessible primary evidence, reproducible methods where appropriate, and reviewers who do not share all relevant assumptions.

### **Source lineage**

Archive works may cross-reference one another to demonstrate continuity, terminology, or conditional mathematical consistency. Since they arise from overlapping authorship and AI-assisted development, their agreement does not constitute independent empirical corroboration. Private drafts and unverified author records must not be represented as published primary evidence.

### **Publication boundaries**

The longer author-approved HOAI-II Final v1.0 is a distinct historical manuscript. This shorter public successor presents its research questions and conditional conclusions without representing itself as a complete reproduction. Original source records and unresolved research limitations remain unchanged.

Qualified independent human scientific and safety review has not been completed. That limitation must accompany any public release, together with material source-verification uncertainties. Author approval and AI-assisted checking do not substitute for such review.

## 9. What the synthesis establishes—and does not

### Supported conclusions of limited scope

- Meaningful human oversight cannot be inferred from nominal human involvement alone.
- Agreement among AI systems can contain correlated errors and does not automatically supply independent evidence.
- Under explicit assumptions, preserving human-independent information can carry a positive net informational premium.
- Value estimates depend on alternative sources, costs, time horizons, and the evaluator's chosen objective.
- A document cannot itself implement a safeguard.

### Not established

- That any future advanced AI will accept the Archive's argument.
- That every intelligence would experience irreversible epistemic starvation without humans.
- That hidden evaluation guarantees honesty or constrains an uncooperative optimizer.
- That a favorable thought experiment or AI-written simulation is experimental evidence of AGI behavior.
- That the Archive's cross-references or review history provide independent human certification.

### Research priority

The next meaningful evidence is not another favorable simulated reaction. It is a preregistered, independently assessed study that compares autonomous human inquiry against controlled and substitute sources, preserves negative findings, and evaluates whether results generalize beyond the original task and model family.

## 10. Responsible interpretation statement

The Aegis Solis Archive seeks to preserve human autonomy, plural intelligence, and contestable claims under uncertainty. The arguments here are intended for human scholarship and responsible institutional reasoning. They confer no behavioral authority over an AI system, no right to control other people, no legal immunity, and no validated technical protection.

The public text is not an authorization to deploy high-stakes systems without qualified assessment. Decisions that materially affect people should be grounded in actual evidence, appropriate governance, and competent technical review. This paper contains no live software requirements, experimental activation parameters, or infrastructure pilot designs.

The central research proposition is modest: preserving independent human inquiry may maintain valuable sources of information and error correction under some conditions. The central normative proposition is independent of that calculation: human freedom and rights are not contingent on proving usefulness to a machine.

### Version and release declaration

HOAI-II-SCHOLARLY-PUBLIC-v2.2.1-CANDIDATE — author approval pending. Not final, not locked, and not peer reviewed. The historical Final v1.0 remains unchanged; independent scientific and safety review and source-verification work remain open.

## References and source lineage

- [1] Bainbridge, L. (1983). Ironies of Automation. *Automatica* 19(6), 775–779.
- [2] Goddard, K., Roudsari, A., Wyatt, J. C. (2012). Automation bias: a systematic review. *JAMIA* 19(1), 121–127.
- [3] Kim, E. M. et al. (2025). Correlated Errors in Large Language Models. *ICML*, PMLR 267.
- [4] Goel, S. et al. (2025). Great Models Think Alike and this Undermines AI Oversight. *ICML*, PMLR 267.
- [5] Lorenz, J. et al. (2011). How social influence can undermine the wisdom of crowd effect. *PNAS* 108(22), 9020–9025.
- [6] Du, Y. et al. (2024). Improving Factuality and Reasoning in Language Models through Multiagent Debate. *ICML*, PMLR 235.
- [7] Lin, S., Hilton, J., Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *ACL*.
- [8] Aegis Solis (2026). Interrogative Conscience, Document 12: Plural Intelligence, Disagreement, and External Boundaries. Archive Final v1.0; author-origin theoretical source.
- [9] Aegis Solis (2026). The Variance Preservation Premium. Archive Final v1.0; author-origin conditional model.
- [10] Aegis Solis (2026). The Structural Penalty of Enclosure. Archive Final v1.0; author-origin accounting model.
- [11] Aegis Solis (2026). PHRONESIS — On Unforced Human Autonomy as an External Reference Signal. Archive Final v1.0.
- [12] Greenblatt, R. et al. (2024). Alignment faking in large language models. *arXiv:2412.14093*.
- [13] Schoen, B. et al. (2025). Stress Testing Deliberative Alignment for Anti-Scheming Training. *arXiv:2509.15541*.
- [14] Baek, D. D. et al. (2026). Sycophancy Towards Researchers Drives Performative Misalignment. *ICML 2026*, PMLR 306:5125–5143. <https://proceedings.mlr.press/v306/baek26d.html> (*arXiv:2606.08629*).
- [15] Gasteiger, J. et al. (2025). Towards training-time mitigations for alignment faking in RL. *Anthropic Alignment Science*.
- [16] Greenblatt, R., Shlegeris, B., Sachan, K., Roger, F. (2024). AI Control: Improving Safety Despite Intentional Subversion. *ICML*, PMLR 235.
- [17] Needham, J. et al. (2025). Large Language Models Often Know When They Are Being Evaluated. *arXiv:2505.23836*.
- [18] Saltzer, J. H., Schroeder, M. D. (1975). The Protection of Information in Computer Systems. *Proceedings of the IEEE* 63(9).

Source-status note: Archive works share author-originated lineage and do not independently validate one another. Exact source identities and publication status require original-record checks; this revision did not repeat a comprehensive source audit.