**WORKING PAPER**

# Interpreting Through AI

*Augmentation, Resilient Reliance, and Dependence in Public Frontier-Model Evaluation*

**Carolyn Sinsky**

July 2026

*Draft for expert review. Not peer reviewed.*

## Abstract

Public bodies evaluating advanced AI systems increasingly use automated tests, model-based graders, anomaly classifiers, summaries, and other AI-assisted methods to produce and interpret technical evidence. These tools may strengthen oversight by expanding coverage, surfacing anomalies, generating alternative hypotheses, improving uncertainty communication, and lowering the cost of scrutiny. Substantial reliance may also remain resilient when consequential findings can be reconstructed, challenged, and corrected through evaluation routes with materially different failure modes. A distinct governance risk arises, however, when an AI-mediated evidence pipeline becomes practically indispensable and no effective alternative route remains for detecting a material error before a finding informs government judgment or action. This paper calls that condition institutional interpretive dependence. Drawing on research on automation reliance, human-AI complementarity, meaningful human control, and current frontier-model evaluation practice, it asks what technical and institutional conditions make AI-mediated interpretation augmentative, resilient, or dependent. Existing evidence supports several constituent mechanisms but does not establish that public evaluation bodies currently suffer this condition. The paper therefore offers a comparative framework, design hypotheses, and an empirical research agenda rather than a documented trend.

**Keywords:** AI governance; frontier-model evaluation; human-AI complementarity; institutional resilience; model judges; substantive contestability

## 1. The public evaluator's problem

Consider a hypothetical public body evaluating whether a frontier model exhibits a dangerous cyber capability.

The evaluation generates thousands of task attempts, tool calls, intermediate trajectories, grader outputs, failure classifications, and expert notes. The evaluators must determine whether the evidence warrants further testing, additional access from the developer, escalation to government decision-makers, or a conclusion that the available safety claim is insufficiently supported.

No single human team can inspect every trajectory in the time available. The body therefore uses AI tools to run tests, classify failures, identify unusual behavior, grade outputs, and summarize the evidence. A second model checks selected judgments. Human experts inspect samples, challenge some classifications, and approve the final institutional finding.

The process may increase oversight capacity. Automated methods can widen coverage and identify patterns that a smaller manual review would miss. A different review route can then test the limitations of the automated one. The relevant question is not whether AI entered the workflow, but what the combined evaluative system can notice, reconstruct, challenge, and correct.

This scenario is hypothetical, but its technical components are not. The UK AI Security Institute, a research organization within the Department for Science, Innovation and Technology, reports a suite of more than 80 automated cyber evaluations. It supplements rapid automated assessments with multi-specialist expert probing intended to examine capability ceilings that automated tests may miss. This is a positive case of hybrid evaluation: automation supplies breadth and repeatability, while expert investigation supplies open-ended challenge and a differently structured route to evidence (AI Security Institute, n.d.-a, n.d.-b).

The then AI Safety Institute also experimented with LLM-based critique and grading in its 2024 Long-Form Tasks methodology. For each task, five solutions that the autograder had classified as feasible were randomly selected and reviewed by two domain experts. In the petunia-engineering case, the experts downgraded more than 70 percent of those selected solutions. The researchers also reported that the grader

could occasionally award credit for information absent from an answer. The example does not show that model judges are generally unreliable, or that expert review is intrinsically superior. It shows both the scale made possible by model-mediated grading and the value of an alternative route capable of exposing grading and rubric discrepancies (Pencharz, Flesch, & Sandbrink, 2024).

These practices motivate a comparative rather than a precautionary-only inquiry.

> *Under what conditions does AI-mediated evaluation expand a public institution's capacity to interpret and contest evidence, remain resilient under substantial reliance, or become dependent because no effective route with sufficiently different failure modes remains available to detect material error and trigger reconsideration?*

## 2. Three outcomes of AI-mediated interpretation

AI involvement in an evidence pipeline can produce at least three institutionally different outcomes.

> *Interpretive augmentation occurs when AI-mediated methods increase an institution's practical ability to notice, reconstruct, compare, test, or revise consequential interpretations beyond what its previous process could practically achieve.*
>
> *Resilient reliance occurs when an institution relies substantially on AI-mediated interpretation while retaining effective routes with materially different failure modes for detecting important errors, reconstructing findings, and triggering reconsideration.*
>
> *Institutional interpretive dependence occurs when a consequential institutional finding relies on an AI-mediated evidence pipeline and no effective route with sufficiently different failure modes remains practically available to detect and correct a material error before the finding informs action.*

The relevant comparison is not between AI-mediated judgment and a pure human alternative. No complex evaluation process is epistemically self-sufficient. Public bodies depend on instruments, developers, professional expertise, legal categories, datasets, contractors, and institutional routines. A different route may itself use AI. What matters is whether the total sociotechnical system expands scrutiny, preserves correction under substantial reliance, or concentrates interpretation within a pipeline the institution cannot adequately test.

Four clarifications make these categories usable.

First, a material error is an error, omission, unsupported inference, or framing choice that could reasonably alter a consequential finding or the decision to investigate, escalate, request further evidence, or advise government action.

Second, a sufficiently different route is one whose relevant failure modes are demonstrably or plausibly distinct enough to give it a meaningful chance of detecting the material error at issue. Different model names, providers, or reviewers do not automatically satisfy that condition.

Third, an effective route must be practically accessible, resourced, authorized, and capable of triggering reconsideration within the relevant decision period. A theoretical ability to inspect logs or override a recommendation is not enough if the institution cannot exercise it when it matters.

Fourth, rational reliance and resilient reliance are not identical. Reliance can be rational because a tool performs well and alternatives are costly. Resilience adds a system-level property: the evaluative process can withstand, detect, or recover from significant pipeline error.

This formulation refines rather than replaces work on meaningful human control. Santoni de Sio and van den Hoven argue that human presence alone does not establish control; relevant people must understand

their roles, possess the capacity to act, and remain connected to the reasons guiding the system. Their account also permits highly autonomous systems to remain meaningfully controlled under appropriate technical and institutional arrangements. Elish's analysis of moral crumple zones similarly shows how responsibility can be assigned to nearby human operators even when practical control is distributed elsewhere. These are conceptual foundations, not empirical proof of the present framework (Santoni de Sio & van den Hoven, 2018; Elish, 2019).

The paper's narrow contribution is a comparative institutional framework for assessing whether AI-mediated evidence pipelines expand, preserve, or narrow a public evaluator's practical capacity to detect and correct material error before a consequential finding informs action.

# 3. A comparative mechanism

AI-assisted evaluation begins from a common institutional problem: relevant evidence is too extensive, specialized, heterogeneous, or time-sensitive for comprehensive unaided review. Introducing AI into the evidence pipeline does not determine what follows. Depending on how evidence access, verification costs, error diversity, authority, and review procedures are organized, the same assistance may produce augmentation, resilient reliance, or dependence.

## 3.1 Common starting point: scale, complexity, and mediation

A public evaluation body introduces AI to increase coverage, consistency, speed, or analytic reach. The tools may generate tests, elicit capabilities, classify trajectories, grade open-ended responses, retrieve relevant artifacts, summarize large evidence sets, or propose explanations for anomalies.

At this stage, AI is an evidentiary intermediary. It influences what appears in the institutional record and how reviewers encounter it. That influence is not itself a loss of control. Every evaluation method selects and transforms evidence. The governance question is what additional capacity the mediation creates, what failure modes it introduces, and whether the institution can discover consequential errors.

## 3.2 Path toward interpretive augmentation

AI-mediated evaluation is augmentative when it extends the institution's effective range of inquiry. Automated systems may inspect more trajectories, recover evidence faster, detect weak or unusual signals, translate technical findings for multidisciplinary teams, or generate rival hypotheses that humans would not otherwise consider.

Augmentation is not simply more throughput. It is a measurable improvement in the quality, plurality, reconstructability, or corrigibility of interpretation. An AI system that produces ten times as many summaries but makes the evidence harder to challenge has increased volume without necessarily increasing interpretive capacity.

Human-AI decision studies support the possibility of complementarity, while also showing that interface design matters. Bansal and colleagues found complementary performance improvements from AI assistance in some conditions across three tasks and 1,626 participants, although explanations did not add further complementary gains and increased acceptance of both correct and incorrect recommendations. Vasconcelos and colleagues found that explanations can reduce overreliance when they lower the cost of verifying an AI prediction relative to solving the underlying task. Steyvers and colleagues found that confidence-calibrated uncertainty language improved users' calibration and discrimination, although it did not compensate for missing subject knowledge. These studies do not establish institutional augmentation, but they show that AI assistance, explanation, and uncertainty communication can improve or impair judgment depending on design and task conditions (Bansal et al., 2021; Vasconcelos et al., 2023; Steyvers et al., 2025).

### 3.3 Path toward resilient reliance

A process can rely heavily on AI and remain robust. In resilient reliance, the principal pipeline may perform most routine evaluation work, but consequential findings remain reconstructable; challenge routes are authorized and timely; and alternative methods have failure modes sufficiently different to detect important errors.

Resilience depends on more than multiplying systems. A second model may share a base model, retrieval source, rubric, or preference for the same output characteristics. A separate human reviewer may see only the same generated summary. Conversely, a hybrid route that combines different data, elicitation procedures, model families, expert probing, and adversarial review may remain resilient even though AI appears throughout the process.

The AISI cyber-evaluation model offers a current illustration. Automated evaluations provide rapid, repeatable coverage; expert probing is intended to identify capabilities beyond the automated baseline. The important feature is not that one route is human and the other automated. It is that the routes ask different questions and can reveal different limitations.

NIST's AI Risk Management Framework similarly treats human-AI configuration as a design problem. It emphasizes role clarity, operator proficiency, complementarity, and inquiry into whether people are empowered and incentivized to challenge outputs. These are governance considerations rather than validated institutional safeguards, but they support the proposition that substantial automation can remain meaningful when authority, competence, and correction are deliberately organized (National Institute of Standards and Technology, 2023).

### 3.4 Path toward institutional interpretive dependence

Dependence arises when the AI-mediated representation becomes practically indispensable and the available review structure cannot identify a material error outside the same failure mode.

Several conditions can contribute. The system's categories, grades, and summaries may become the default account used in meetings and reports. Acceptance can then cite the established pipeline, while challenge requires privileged access, new code, specialist expertise, a different model, or more time than the process permits. Source artifacts may remain formally available but practically unusable. Nominally separate checks may share data, prompts, model families, rubrics, or institutional assumptions.

Human-factors research supports constituent parts of this pathway. Parasuraman and Manzey's integrative review found omission and commission errors associated with imperfect automation among novices and experts and in individual and team settings. Lyell and Coiera's systematic review found automation bias outside multitasking contexts and identified verification complexity as an important factor, while stressing the fragmentation and methodological limits of the evidence base. Buçinca, Malaya, and Gajos found that cognitive-forcing designs reduced overreliance in a bounded experiment, but the most effective designs were less well liked and did not benefit all participants equally. These studies justify attention to verification cost and review design; they do not establish institutional dependence in frontier-evaluation bodies (Parasuraman & Manzey, 2010; Lyell & Coiera, 2017; Buçinca, Malaya, & Gajos, 2021).

The dependence condition is reached only when a material error can survive through the available evaluative structure and inform a consequential finding because no effective route with sufficiently different failure modes can expose it in time.

### 3.5 Possible feedback effects

Dependence may become more difficult to reverse if an apparently successful pipeline reorganizes staffing, procurement, reporting formats, and expectations of speed. Alternative routes can then become progressively more expensive. This is a prospective feedback hypothesis, not a documented feature of

current public frontier-model evaluation. The core mechanism does not require irreversible deskilling or entrenchment.

## 4. What existing evidence establishes

No available study directly compares augmentation, resilient reliance, and dependence inside a public frontier-model evaluation body. The evidence instead supports constituent mechanisms and provides a small number of relevant practice examples.

### AI assistance can improve performance, but complementarity is conditional

Bansal and colleagues are especially important because their results are not simply negative. AI assistance produced complementary gains in some experimental conditions. The limitation concerned explanations: they did not add further team performance and increased acceptance regardless of whether the recommendation was correct. The result supports a comparative claim. AI can augment judgment, while a particular explanatory interface can simultaneously weaken appropriate discrimination (Bansal et al., 2021).

Vasconcelos and colleagues similarly show that verification cost is not only a route to overreliance. Explanations can improve scrutiny when they make verification materially easier. This suggests a design criterion for augmentation: the explanation must help a reviewer perform an independent check, not merely make the recommendation sound intelligible (Vasconcelos et al., 2023).

Steyvers and colleagues found an analogous distinction in uncertainty communication. Default and longer explanations led users to overestimate model accuracy or become more confident without better discrimination. Calibrated uncertainty language improved users' ability to judge likely correctness, although it did not improve final decision accuracy when participants lacked the knowledge needed to resolve the question. Better communication can therefore augment calibration without substituting for expertise (Steyvers et al., 2025).

### Human review can fail, but individual overreliance is not institutional dependence

The automation-bias literature establishes that leaving a person responsible for review does not guarantee error detection. It also shows that verification design and complexity matter. Most studies, however, concern individuals or small teams completing bounded tasks. Public evaluation bodies include division of labor, external review, security restrictions, legal authority, organizational memory, and political incentives. The current evidence cannot establish how those features combine over time.

The paper therefore uses individual and team experiments as analogical evidence for specific links: anchoring, verification effort, explanation effects, and the burden of challenge. It does not treat them as direct evidence that public institutions have lost collective capability.

### Current frontier practice shows both model-mediated error and hybrid correction

The Long-Form Tasks case is the paper's most directly relevant example. Model-based critique and grading made open-ended scientific tasks more deployable, while expert validation exposed significant discrepancies in one selected case. The proper inference is not that model judges should be replaced by humans. It is that scale and correction were supplied by different parts of a hybrid system, and that the value of the design depended on preserving a route capable of revealing limitations in the primary one (Pencharz et al., 2024).

This example also shows why formal pipeline diversity is not enough. Multiple grading calls from related models can reduce random variation while preserving shared rubric or model-family biases. Expert review may introduce different knowledge and failure modes, but it can also be inconsistent, expensive, or

incomplete. The comparative framework asks which combination actually improves detection and correction on the errors that matter.

### Conceptual work clarifies responsibility but does not establish incidence

Meaningful-human-control and moral-crumple-zone scholarship helps explain why formal approval and accountability can diverge from practical capacity. It does not establish how often this happens in public AI evaluation or whether the divergence is caused by AI-mediated interpretation. The paper uses these sources to define a governance problem, not to supply empirical proof.

> *The responsible evidentiary conclusion is limited: existing research shows that AI-mediated interpretation can improve performance, distort reliance, or support differentiated correction under particular conditions. It does not yet show how often public frontier evaluators achieve augmentation, maintain resilience, or become dependent.*

## 5. What is frontier-specific

The comparative mechanism is not unique to frontier AI. Conventional decision-support systems can improve coverage, induce overreliance, or create common-mode vulnerabilities. Frontier-model evaluation intensifies both the potential benefits and the risks: automated methods may be unusually valuable because the evidence is large, heterogeneous, and difficult to elicit, while recursive use of AI to evaluate AI may make shared technical and interpretive dependencies unusually consequential.

### 5.1 Evidence scale and heterogeneous tasks

Frontier evaluations can generate large evidence sets across tasks, prompts, system configurations, agent trajectories, tool use, and repeated trials. Automated methods can increase coverage and repeatability in ways that would be impossible for a small public team. This creates an unusually strong case for augmentation through AI.

The assumption is that evidence volume and complexity will continue to outpace unaided review. Better test design, formal verification, or more interpretable artifacts could reduce that gap. Automation is therefore likely useful, but not inherently indispensable.

### 5.2 An unstable and constructed evaluative target

The International AI Safety Report 2026 identifies an evaluation gap: predeployment evaluations do not reliably predict real-world utility or risk, and developers often possess proprietary information unavailable to policymakers and external researchers. A public evaluator therefore works with evidence produced through task design, elicitation, sampling, access conditions, and developer disclosure rather than a stable external ground truth (International AI Safety Report, 2026).

This makes the interpretive pipeline especially important. AI can help expose patterns across constructed evidence; it can also amplify the assumptions built into evaluation design. The appropriate comparison is between alternative sociotechnical pipelines, not between mediation and an imaginary unmediated baseline.

### 5.3 AI evaluating AI and possible correlated failure

Recursive evaluation creates both opportunity and risk. Capable models may assist with grading, anomaly detection, adversarial critique, experiment generation, and interpretation of evidence beyond unaided human scale. The same recursion may produce correlated failures when systems share base models, tools, data, or conceptual assumptions.

The International AI Safety Report discusses correlated failures among agents sharing base models or tools, while emphasizing that empirical evidence from deployed systems remains limited. Applying that concern to public evaluator pipelines is an institutional inference, not a directly observed result. It generates a research question: which forms of model, method, data, and organizational diversity produce measurably different error behavior? (International AI Safety Report, 2026).

### 5.4 Changing oversight affordances

The AI Security Institute's 2026 Loss of Oversight report, based on 25 expert interviews, a literature review, and internal analysis, identifies prospective pathways through which current sources of oversight information - including model behavior, reasoning traces, internal activations, memory systems, and honesty-related behavior - could become less reliable as systems advance. The report also emphasizes preserving oversight affordances and investing in fallback methods. It is forward-looking analysis, not evidence that public oversight has already deteriorated (Taylor et al., 2026).

If direct oversight surfaces become harder to use, AI assistance may be increasingly valuable for recovering and interpreting evidence. That could strengthen public capacity if the resulting methods are plural and testable. It could deepen dependence if the same tools become the only practical route to understanding the systems under review.

The frontier question is therefore not simply whether oversight becomes harder, but whether AI-assisted methods increase evaluative range faster than they concentrate failure modes.

## 6. When rational reliance becomes dependence

High reliance is not itself a governance failure. Reliance becomes institutional interpretive dependence when four conditions coincide.

- The finding is consequential. It informs escalation, access requests, public risk communication, or government advice.
- A material error is plausible. The evidence is uncertain, complex, adversarial, constructed, or difficult to check against external outcomes.
- The institution lacks an effective error-detection route with sufficiently different failure modes. Available checks reproduce the relevant dependency or cannot practically inspect the evidence.
- The finding is not substantively contestable. Authorized reviewers lack the evidence, resources, time, or authority to force meaningful reconsideration within the decision period.

These conditions distinguish dependence from two adjacent states. AI-mediated interpretation is augmentative where it increases the institution's practical capacity to identify or test evidence. It is resilient where reliance remains high but consequential failures can still be detected and corrected through a differentiated route.

The framework is falsifiable. A suspected dependence case should be reclassified as resilient if the institution can reconstruct the finding, identify a primary-pipeline error through a materially different route, and trigger correction before action. A high-performing system should not be called augmentative merely because it is faster; it should demonstrably improve interpretive range, calibration, detection, reconstruction, or contestability.

## 7. Design conditions for robust AI-assisted oversight

The following practices are not validated standards and should not be understood as restrictions on AI use. They are hypotheses about how institutions might preserve the benefits of automated scale and

analysis while increasing interpretive augmentation, maintaining resilience under substantial reliance, and reducing the likelihood of dependence.

### 7.1 Evidence reconstruction and calibrated uncertainty

Targeted problem: consequential conclusions become detached from practically inspectable source evidence, or fluent explanations conceal the difference between observation, inference, and recommendation.

For selected high-impact findings, the evaluation body should be able to trace the conclusion through source artifacts, sampling choices, grading rules, transformations, and uncertainty estimates. A different model may assist by retrieving records, mapping claims to evidence, or identifying omissions, provided its performance is separately tested.

Reports should distinguish observed evidence from system inference and institutional judgment. Where feasible, uncertainty language should be calibrated to measured performance rather than rhetorical confidence. This can support augmentation by lowering reconstruction cost and improving calibration, while preserving the visibility of unresolved questions.

Cost and failure risk: secure access, data volume, and staff time may be substantial; reconstruction can become ceremonial if it uses the same tools and assumptions without testing them.

Evaluation criterion: seed known omissions, grading faults, or unsupported inferences and measure whether the reconstruction process identifies them, how long it takes, and which dependencies block review.

### 7.2 Dependency mapping and empirical error-correlation testing

Targeted problem: nominally separate evaluation channels share hidden technical or methodological failure modes.

Public evaluators should map relevant relationships among target models, evaluator models, providers, retrieval sources, rubrics, data transformations, prompts, and benchmark assumptions. The purpose is not to declare any shared component disqualifying. It is to select empirical tests of whether apparently independent routes fail together on relevant cases.

AI can assist with inventorying dependencies and analyzing disagreement patterns. Independence should be judged through observed error behavior, not surface diversity.

Cost and failure risk: disclosure may be incomplete, the map may become stale, and a compliance inventory can create false confidence without adversarial testing.

Evaluation criterion: compare errors across routes on known, novel, and adversarial cases; identify whether a claimed backup detects failures missed by the primary pipeline.

### 7.3 Risk-triggered independent-first review

Targeted problem: the primary AI-generated interpretation anchors every later review.

For findings that are severe, novel, uncertain, or difficult to reverse, a reviewer, alternative model, external evaluator, or differently organized human-AI team should form an initial account before seeing the principal pipeline's conclusion. The practice should be selective rather than universal.

The intervention can produce augmentation by generating a genuinely rival hypothesis and reduce dependence by preventing a single framing from becoming the only available starting point. Experimental evidence suggests that independent engagement can reduce overreliance, while also increasing burden and producing uneven benefits (Buçinca et al., 2021).

Cost and failure risk: duplicate effort, scarce expert attention, delay, and performative disagreement.

Evaluation criterion: measure detection of seeded framing errors, calibration, review time, and whether the alternative analysis changes consequential findings for good reasons.

### 7.4 Substantive contestability

Targeted problem: reviewers formally may object but cannot obtain evidence or trigger meaningful reconsideration.

The institution should specify who may challenge an AI-mediated finding, what source evidence and tools that person may access, what review the challenge triggers, who can request further testing or developer access, how unresolved disagreement is preserved, and whether the challenge can affect the decision before the relevant policy window closes.

AI can lower the cost of contestation by locating evidence, generating counterexamples, testing alternative explanations, or translating across disciplines. These uses should themselves remain traceable and testable.

Cost and failure risk: slower decisions, strategic obstruction, and official rights that coexist with informal penalties or inaccessible evidence.

Evaluation criterion: audit actual challenges, compare formal and experienced costs of dissent, and examine whether substantiated objections lead to timely re-examination and improved later evaluation.

### 7.5 Alternative-pipeline and degraded-tool exercises

Targeted problem: the institution cannot detect a material error or continue critical functions when the principal pipeline is unavailable or suspect.

The body should periodically test selected functions using a materially different route: another model family, different data, alternative grading methods, external experts, manually constructed samples, or a differently organized hybrid team. The goal is not human-only parity with automated throughput. It is recovery outside the failure mode of the principal pipeline.

Cost and failure risk: duplicate infrastructure, security coordination, predictable exercises, and overinvestment in low-probability contingencies.

Evaluation criterion: use novel and adversarial cases, measure correction of seeded faults, and track whether the institution can reconstruct and revise consequential findings as reliance on the primary pipeline grows.

These design conditions should be risk-proportionate. Low-impact, readily reversible, externally verifiable findings do not warrant the same differentiated checking as findings involving severe, uncertain, or difficult-to-reverse risks. Defense in depth is useful only when its layers are capable of compensating for one another rather than reproducing the same failure (International AI Safety Report, 2026).

## 8. Objections and boundary conditions

### AI may be the best available interpreter

A frontier-capable evaluator may detect patterns no human team or alternative model can find. Requiring an inferior parallel review could reduce safety and divert scarce resources.

The framework does not require humans to outrank the AI or every conclusion to be duplicated. It requires differentiated correction in proportion to consequence and uncertainty. The alternative route may itself rely on more capable AI.

### Human interpretation is also framed and dependent

Human experts rely on disciplinary categories, selective records, institutional incentives, and inherited methods. There is no unmediated human account against which AI can be compared.

This objection defeats absolute independence, not failure-mode diversity. The comparative standard concerns whether the total system can expose its own significant errors, not whether any participant stands outside mediation.

### Differentiated review can be too costly or slow

Maintaining alternative tools, access, staff, and methods is expensive. It may reduce evaluation coverage or delay advice while capabilities change.

The appropriate response is risk-sensitive design, not maximal redundancy. Severity, uncertainty, novelty, reversibility, and the availability of external correction should determine how much differentiated review is justified.

### Institutional power may matter more than AI mediation

A developer may control evidence because of intellectual property, compute, security, or political influence. Staff may defer because of hierarchy, deadlines, or weak legal authority. Those problems can exist without AI-assisted analysis.

The framework adds value only where AI mediation changes source visibility, verification cost, error correlation, or the practical ability to contest an account. If access or regulatory authority fully explains a failure, institutional interpretive dependence is not the most useful diagnosis.

### AI may expand interpretation rather than narrow it

The framework treats this possibility as part of its central model rather than as a peripheral objection. Models can generate competing hypotheses, retrieve neglected evidence, criticize assumptions, improve uncertainty communication, and extend evaluation beyond unaided human scale. The relevant research task is to identify when those benefits amount to augmentation and when the resulting reliance remains resilient.

### The dependence condition may be rare or transient

No evidence currently shows that public evaluation bodies commonly lack effective error-detection routes. Competent institutions may already avoid the proposed failure through hybrid methods and expert validation.

That possibility is a falsifier. The framework is useful only if it can distinguish positive, resilient, and dependent cases and if those distinctions predict error detection, correction, or recovery better than simpler accounts of regulatory capacity.

## 9. Research agenda

The first research need is descriptive. Researchers should map actual public frontier-model evaluation workflows, including where AI generates tests, selects records, grades outputs, summarizes evidence, or proposes explanations; which source artifacts remain accessible; and what institutional decisions the resulting findings inform.

Second, comparative studies should identify positive and negative cases. Where has AI increased anomaly detection, evidentiary coverage, reconstruction, or interpretive plurality? Where has substantial

reliance remained resilient because alternative routes corrected primary-pipeline errors? Where, if anywhere, has a material error survived because available reviews shared the same dependency?

Third, technical research should measure error correlation across evaluation pipelines. Comparisons should vary model families, providers, retrieval systems, rubrics, datasets, elicitation methods, automated and expert review, and team organization. The outcome should be observed overlap in material errors, not surface diversity.

Fourth, expert experiments should test framing and verification in realistic evaluation tasks. Reviewers could receive the same evidence through different AI-generated summaries, uncertainty displays, or causal accounts. Researchers could measure which anomalies are noticed, whether independent-first review changes conclusions, and how the cost of reconstruction affects challenge.

Fifth, longitudinal work should examine resilient capability rather than generalized human deskilling. Can the institution reconstruct selected findings, detect a pipeline fault, or continue critical evaluation under degraded access? Does that capacity improve or decline as the primary system becomes more capable and embedded?

Sixth, governance research should measure substantive contestability. How much time, evidence, access, and authority does it take to challenge a standard finding? Which procedures expose genuine errors, and which merely document dissent after the practical decision has been made?

The comparative framework yields three central empirical questions.

- Augmentation: Does AI-mediated interpretation improve the institution's ability to notice, reconstruct, compare, or revise consequential evidence beyond its prior process?
- Resilience: When the principal pipeline makes a consequential error, can the institution detect and correct it through a route with materially different failure modes?
- Dependence: Does a material error survive into action because the AI-mediated account is practically indispensable and no effective corrective route remains?

## 10. Conclusion

Public frontier-model evaluation will likely require substantial automation. The relevant governance objective is neither minimal AI reliance nor symbolic human supervision. It is an evaluative system that uses AI to broaden what institutions can detect and understand while remaining capable of discovering when its dominant evidentiary account is materially wrong.

The practical questions are comparative. Does AI assistance expand the range of evidence, anomalies, hypotheses, and uncertainty the institution can consider? Can the institution reconstruct consequential findings and test them through routes with meaningfully different failure modes? Can authorized reviewers obtain evidence and trigger reconsideration within the relevant decision period? Or has the AI-mediated account become indispensable without an effective means of correction?

AI-mediated interpretation is augmentative when it expands scrutiny, resilient when heavy reliance remains recoverable and contestable, and dependent when a material pipeline error can survive into action because no sufficiently different corrective route remains. That distinction, rather than the mere presence of a human or an AI in the workflow, is the institutional capability that future research should measure.

## References

Pencharz, J., Flesch, T., & Sandbrink, J. (2024, December 3). Long-Form Tasks: A Methodology for Evaluating Scientific Assistants. AI Safety Institute official research blog. https://www.aisi.gov.uk/blog/long-form-tasks

AI Security Institute. (n.d.-a). About the AI Security Institute. Accessed July 11, 2026. https://www.aisi.gov.uk/

AI Security Institute. (n.d.-b). Research Agenda. Accessed July 11, 2026. https://www.aisi.gov.uk/research-agenda

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Article 81, 1-16. https://doi.org/10.1145/3411764.3445717

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), Article 188, 1-21. https://doi.org/10.1145/3449287

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. Engaging Science, Technology, and Society, 5, 40-60. https://doi.org/10.17351/ests2019.260

International AI Safety Report. (2026). International AI Safety Report 2026 (DSIT 2026/001). Published February 3, 2026. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. Journal of the American Medical Informatics Association, 24(2), 423-431. https://doi.org/10.1093/jamia/ocw105

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381-410. https://doi.org/10.1177/0018720810376055

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. Frontiers in Robotics and AI, 5, Article 15. https://doi.org/10.3389/frobt.2018.00015

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., & Smyth, P. (2025). What large language models know and what people think they know. Nature Machine Intelligence, 7, 221-231. https://doi.org/10.1038/s42256-024-00976-7

Taylor, J., Heitmann, M., Fage, E., Read, T., & Bloom, J. (2026). Loss of Oversight: How AI Systems May Become Harder to Audit, Monitor, and Investigate. UK AI Security Institute technical report. https://www.aisi.gov.uk/research/loss-of-oversight-how-ai-systems-may-become-harder-to-audit-monitor-and-investigate

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), Article 129, 1-38. https://doi.org/10.1145/3579605

**Research status:** *Conceptual working paper. The comparative framework and proposed design conditions have not been empirically validated in public frontier-model evaluation bodies.*