Human oversight of AI

What makes review effective and what evidence decision makers should require

Research by The Actual Factual | Brief 003 | Edition 1 | 14 November 2026

Purpose and scope

A focused evidence review for legislators, boards, schools and public agencies. It examines selected experiments, NIST and GAO frameworks, and the original 2024 EU AI Act text. It does not assess a named deployment, certify compliance, or establish a universal staffing ratio. Reviewed 14 November 2026.

Executive summary

A person approving an AI recommendation does not establish that mistakes will be detected. The reviewed experiments support testing the decision process rather than assuming review improves it. [3] [6]

Oversight needs an accountable role and a workable route to intervene. NIST connects risk management with responsibilities and training. A governance commitment is not evidence of effective implementation. [1]

Our recommendation is to evaluate the complete decision process: the evidence reviewers see, their actual decisions, missed errors, justified overrides and consequences for affected people. GAO's framework provides a structure for asking about governance, data, performance and monitoring. [8]

Begin with the decision that people are making

A school reviewing an AI-generated lesson, a public agency assessing an application and a board approving an automated process face different decisions. Our proposed starting point is a written description of the decision, who may be affected, which part the AI performs and who can change the result. These are illustrative contexts, not findings about any named institution.

NIST AI RMF 1.0 is a voluntary framework. Its GOVERN provisions distinguish operational responsibilities, staff training and executive accountability; Appendix C discusses the varied results of human-AI interaction. Assigning a reviewer therefore raises a further research question: what resources and evidence make that role effective here? [1]

Acceptance and accuracy are different outcomes

Sele and Chugunova's 2024 online prediction experiment involved 292 participants. Allowing adjustment made algorithmic delegation more attractive, but the human-in-the-loop design reduced decision accuracy in that experiment. Participants were less likely to intervene on the least accurate recommendations. This supports scrutiny of review design; it does not show that all human oversight is harmful. [3]

Bansal and colleagues tested three tasks with AI accuracy comparable to human performance. Combined performance improved, but the tested explanations added no further improvement and increased acceptance of incorrect advice. Confidence in an explanation should not substitute for measured outcomes. [4]

The evidence does not support a single bias story

Alon-Barkat and Busuioc's three Dutch experiments did not find a general tendency to follow algorithmic advice more than equivalent human-expert advice. One citizen study found greater adherence when advice matched group stereotypes; the civil-servant study did not reproduce that selective-adherence pattern. This counterevidence matters: automatic deference is a risk to test, not a universal description of public employees. [5]

Agudo and colleagues manipulated when participants saw erroneous automated support in simulated judgments about defendants. Incorrect advice was particularly damaging when presented before participants formed their own judgments. These were experimental tasks, not an audit of courts or a demonstration that one interface sequence is always safest. [6]

An intervention can have costs as well as benefits

Buçinca and colleagues tested interventions designed to make participants think before relying on AI advice. In their 199-person experiment, these interventions reduced overreliance relative to simpler explainable-AI approaches. The strongest designs received less favorable subjective ratings, and benefits differed with participants' motivation for effortful thinking. A promising intervention still needs evaluation in its intended setting. [7]

Our practical inference is to test review under realistic workload, information access and time constraints before treating it as a safeguard. Record when reviewers disagree correctly, when they introduce errors and when the process delays a needed decision. These are proposed evaluation questions, not measured outcomes from a deployment reviewed here.

Separate written requirements from demonstrated performance

Articles 14 and 26 of the original 2024 EU AI Act connect high-risk AI oversight with proportionality, interpretation, intervention and supported personnel. This is a historical governance reference. We do not determine current applicability, subsequent amendments, which organizations are covered or when duties apply. [2]

GAO's 2021 framework includes assessment questions and audit procedures across four principles. A useful next step is to ask for the records behind an assurance of oversight: the assigned responsibility, evaluation plan, monitoring results and corrective action. A policy can establish an organizational commitment; performance records are needed to examine whether it is fulfilled. The latter is our analytical distinction. [8]

Evidence record

The table identifies the record, the finding it supports, and the limit of that evidence. Numbered citations open the original source.

Date or record	Evidence and limits
2021	**Team performance experiment.** Human-AI combinations improved performance; the tested explanations added no further improvement. [4] **Limit:** Three tasks, not a universal result.
2021	**Review design experiment.** Buçinca and colleagues found lower overreliance with cognitive-forcing interventions. [7] **Limit:** Benefits came with usability and participant-difference tradeoffs.
30 June 2021	**Accountability framework.** GAO organized assessment around governance, data, performance and monitoring. [8] **Limit:** A framework is neither a certification nor an evaluation of the reader's system.
8 February 2022	**Public-sector experiments.** Alon-Barkat and Busuioc did not find general automation bias relative to equivalent human advice. [5] **Limit:** Dutch experimental settings; results differ across samples and measures.
January 2023	**Voluntary risk framework.** NIST AI RMF 1.0 specifies responsibilities and training for risk management. [1] **Limit:** Adoption does not establish successful implementation.
7 January and 9 February 2024	**Timing and adjustment experiments.** Agudo and colleagues studied advice timing; Sele and Chugunova studied adjustable recommendations. Both identified conditions in which human decisions deteriorated. [6] [3] **Limit:** Different experimental designs cannot supply a single real-world failure rate.
12 July 2024 publication	**Original EU AI Act.** Articles 14 and 26 describe oversight capabilities and personnel requirements. [2] **Limit:** Historical text; subsequent amendments, applicability and transition rules require separate legal analysis.

Claims findings and unknowns

Claim 1

Claim under examination: A human signs off, so the AI is adequately supervised.

Finding: The signature describes a process step. The experimental evidence does not establish that such a step reliably improves accuracy. [3] [6]

What remains unknown: Whether reviewers detect the errors that matter in this deployment.

Claim 2

Claim under examination: An explanation makes the recommendation safer.

Finding: The tested explanations increased acceptance of incorrect advice. [4]

What remains unknown: Whether the proposed explanation improves decisions here.

Claim 3

Claim under examination: People inevitably defer to algorithms.

Finding: The Dutch studies provide counterevidence to that generalization. [5]

What remains unknown: How this workforce responds over repeated use, under its actual incentives.

Questions for decision makers

These are recommended oversight questions. They identify records that could support an answer and do not presume a particular conclusion.

Question 1

Who can pause, override or reject the AI-supported decision?

Why this matters: Our recommendation is to demonstrate practical authority rather than merely identify a job title. [1]

Who can answer: Operational owner and accountable executive.

Records to request: Role descriptions, access permissions, escalation procedure and a demonstrated intervention.

Question 2

Does the combined process outperform each baseline?

Why this matters: Our recommendation is to test added value. [4]

Who can answer: Evaluation lead and independent assessor.

Records to request: Protocol, representative cases, human-only and AI-only results, error severity and uncertainty.

Question 3

Can reviewers form an independent judgment before seeing a recommendation?

Why this matters: Timing changed susceptibility to incorrect advice in an experiment; its relevance here should be tested. [6]

Who can answer: Service designer and operational supervisor.

Records to request: Interface sequence, alternative-workflow test, reviewer reasoning and case-level outcomes.

Question 4

Which errors reach people despite review?

Why this matters: Our proposed monitoring approach examines missed errors and consequences, not only how often staff press an override button. [8]

Who can answer: Quality lead, complaints owner and internal audit.

Records to request: Incident log, error sampling plan, appeal outcomes, corrective-action record and privacy protections.

Question 5

Does review remain effective when demand rises or staff change?

Why this matters: Interventions have tradeoffs. Our recommendation is to evaluate workload and accessibility, then specify conditions that require redesign or suspension. [7]

Who can answer: Service manager and staff responsible for review.

Records to request: Staffing assumptions, training assessments, timed exercises, accessibility testing and review capacity plan.

Connections and their limits

These relationships are a reading aid. A connection does not establish causation, strength of influence, or motive. The interactive version is available on the report landing page.

Reviewer authority and Ability to intervene

Documented relationship: The reviewed EU provisions connect human oversight with authority and intervention capabilities. [2]

Limit: The legal text does not establish who has working permissions in a particular system.

Explanations and Acceptance of advice

Documented relationship: The experiments linked explanations with acceptance, including wrong advice. [4]

Limit: Bounded by the tested tasks and designs.

Advice timing and Independent judgment

Documented relationship: Agudo's experiments varied when incorrect support was shown and found different judgment outcomes. [6]

Limit: An experimental effect is a reason to test workflow, not a universal design prescription.

Review friction and Error detection

Documented relationship: Cognitive-forcing designs reduced overreliance in Buçinca's study while receiving less favorable ratings. [7]

Limit: More friction is not automatically better; effects were context-dependent.

Monitoring records and Accountability

Editorial analysis: Our inference from GAO's assessment structure is that traceable records make oversight claims examinable. [8]

Limit: Records enable scrutiny; their existence does not prove that harm was prevented.

Method and corrections history

Sources were selected for original empirical evidence or authoritative governance text, including findings that challenge a simple pro- or anti-automation account. This is a focused review, not a systematic review or meta-analysis. Study outcomes remain attached to their tasks and samples. Proposed questions and cross-source inferences are labelled as our recommendations or analysis. We did not audit an operational system, measure staff performance, or determine legal compliance. Corrections should identify the affected claim, source and revision date.

14 November 2026 | Edition 1. Initial publication of this research brief. No corrections recorded in this edition. Future substantive updates should identify what changed and the supporting evidence.

Web report and future updates

Source register

Each numbered citation and source title links to the original record. Source dates identify the cited version, not necessarily its current legal or operational status.

[1] Artificial Intelligence Risk Management Framework AI RMF 1.0

National Institute of Standards and Technology | January 2023

Primary voluntary framework. GOVERN 2 and 3.2; Appendix C. Source version is 1.0, not an assertion that it is the newest guidance.

[2] Regulation EU 2024 1689 original act Articles 14 and 26

Official Journal of the European Union via EUR-Lex | 12 July 2024 publication

Original enacted text. Article 14 on printed pages 60-61 and Article 26 beginning on printed page 67. No current-applicability assessment is made.

[3] Putting a human in the loop Increasing uptake but decreasing accuracy of automated decision making

Daniela Sele and Marina Chugunova in PLOS ONE | 9 February 2024

Original experiment, 292 participants. Findings concern the tested prediction task.

[4] Does the Whole Exceed its Parts The Effect of AI Explanations on Complementary Team Performance

Gagan Bansal and colleagues at CHI | May 2021

Original research on three tasks; corporate-hosted publication page.

[5] Human AI Interactions in Public Sector Decision Making Automation Bias and Selective Adherence to Algorithmic Advice

Saar Alon-Barkat and Madalina Busuioc in JPART | 8 February 2022 online publication; January 2023 issue

Original study. Three Dutch experiments, including a civil-servant sample. Includes findings that qualify a simple automation-bias narrative.

[6] The impact of AI errors in a human in the loop process

Ujué Agudo and colleagues in Cognitive Research Principles and Implications | 7 January 2024

Original two-experiment study using simulated judgments; the second experiment was preregistered.

[7] To Trust or to Think Cognitive Forcing Functions Can Reduce Overreliance on AI in AI assisted Decision making

Zana Buçinca Maja Barbara Malaya and Krzysztof Z Gajos | 19 February 2021 preprint; 2021 CSCW publication

Authors' research manuscript; related publication DOI 10.1145/3449287. Experiment with 199 participants.

[8] Artificial Intelligence An Accountability Framework for Federal Agencies and Other Entities

U.S. Government Accountability Office | 30 June 2021

Primary framework GAO-21-519SP. Supports evaluation and audit questions, not a claim about a particular institution's performance.