



2026

Human in the Loop

BRITE INSTITUTE

The Function of Human Oversight

Human oversight is positioned as the control of last resort: the mechanism intended to detect what every other safeguard has failed to prevent. It is load-bearing by design.

The underlying logic is straightforward. An AI system will sometimes produce incorrect output. It may be biased, confidently mistaken, or operating beyond the conditions for which it was designed. A qualified person is therefore placed between the system's output and its consequence. The clinician reviews the recommendation; the loan officer examines the decision; the judge weighs the risk score. Where the system errs, the reviewer is expected to intervene.

A considerable burden rests on this arrangement. Human oversight is expected to detect:

- **Erroneous output.** Incorrect results that pre-deployment testing did not anticipate.
- **Bias.** Systematic differences in performance across demographic groups that aggregate accuracy metrics conceal.
- **Edge cases.** The atypical patient, file, or circumstance absent from the training distribution.
- **Distributional drift.** The gradual degradation of a model deployed into a population that no longer resembles the one from which it learned.
- **Accountability.** The requirement that a person, rather than a system, remains answerable for the outcome.

These expectations are codified in the EU Artificial Intelligence Act. Article 14 requires that high-risk AI systems be designed so that they can be effectively overseen by natural persons, and provides that oversight shall aim to prevent or minimize risks to health, safety or fundamental rights – particularly where such risks persist despite the application of other requirements.

The EU Regulatory Assumption

On August 2, 2026, Article 14 of the EU Artificial Intelligence Act will begin enforcing human-in-the-loop oversight for high-risk systems such as education, employment, and law enforcement.

Within the same article, the act acknowledges the principal threat to the control it establishes: human failure. Article 14(4)(b) requires that persons assigned to oversight be enabled to remain aware of “the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias)”, specifically with respect to systems providing information or recommendations for decisions taken by natural persons.

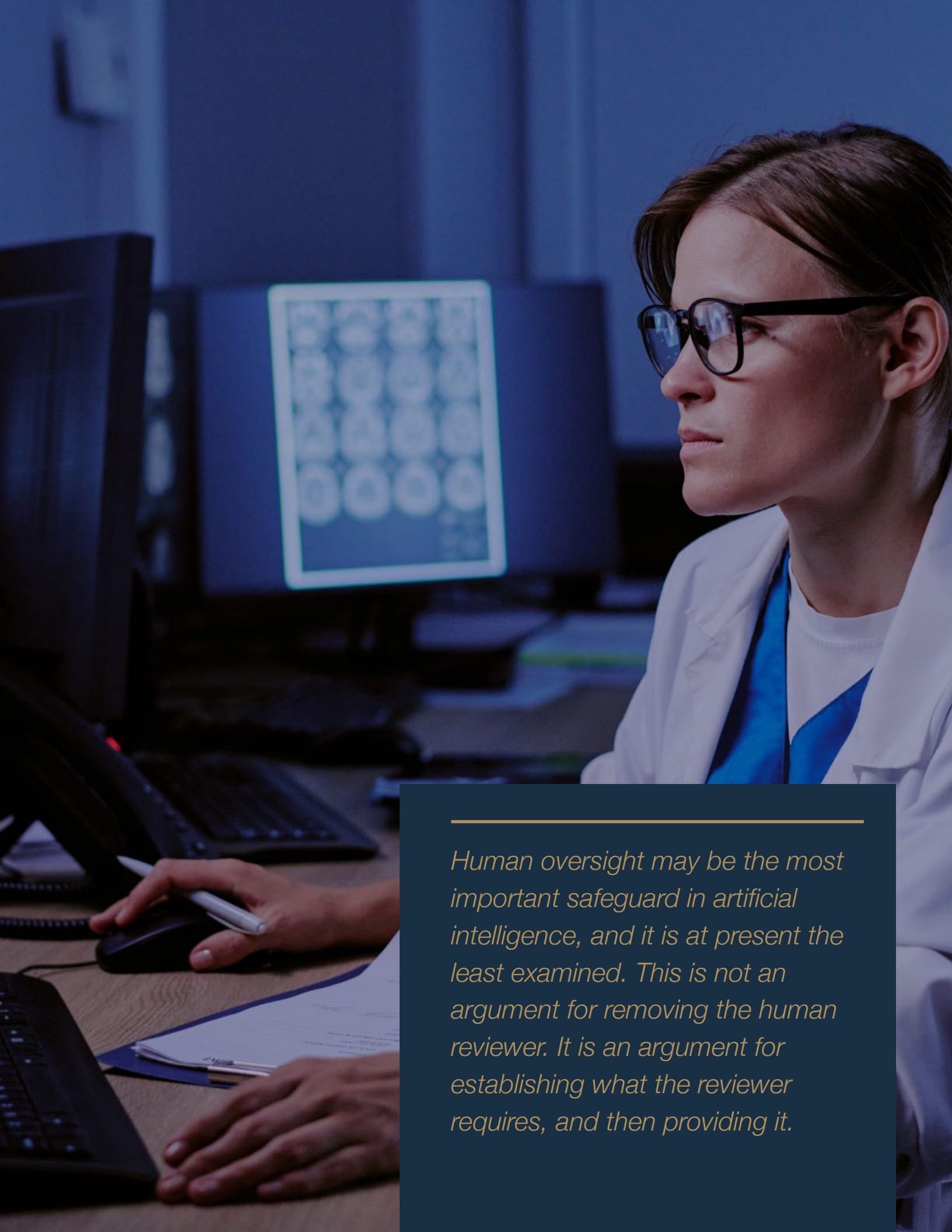
The Regulation therefore identifies a known failure mode of its own primary safeguard, and specifies awareness as the mitigation.



Simply being aware of human fallibility does not, in general, remove said fallibility.

Further measures are needed to make sure the human-in-the-loop catches and appropriately responds to AI error.

BRITE Institute argue that awareness is not a control. Making the human-in-the-loop aware that they are subject to cognitive limitations does not ameliorate those cognitive limitations. To make human-in-the-loop an effective risk management method, it is critical further action is taken.



Human oversight may be the most important safeguard in artificial intelligence, and it is at present the least examined. This is not an argument for removing the human reviewer. It is an argument for establishing what the reviewer requires, and then providing it.

The Research

This body of research is sometimes characterized as early or emerging. It is neither. The human factors science concerning the supervision of automation is mature; it is the application of that science to AI oversight that remains untested.

Six findings are of particular consequence.

1. Automation bias affects experts as well as novices

Automation bias – the tendency to over-rely on automated output and to under-detect its failures – is among the most extensively replicated findings in the human factors literature. Goddard and colleagues synthesized cross-domain evidence of over-reliance on decision support as early as 2011, including a finding of direct relevance to any safeguard design: users frequently fail to recognize errors introduced by the automation itself.

A 2025 systematic review by Romeo and Conti, covering 35 peer-reviewed studies published between 2015 and 2025, found over-reliance to be shaped by interacting user, task, and interface factors rather than by individual inattention. A 2026 preregistered randomized experiment by Goulas and colleagues examined the question in isolation: domain experts were presented with identical errors, labeled in some conditions as AI-generated and in others as human-generated, and their correction behavior was compared.

Correction rates differed by label. Attribution to an AI system reduced the scrutiny applied to the same underlying error.

2. Verification is the mediating variable

This finding is among the most consequential in the literature and among the least reflected in practice.

Hondrich and Ruschemeier argue that verification behavior – whether a reviewer in fact checks the output – is the central practical mediator of automation bias. Their conclusion is unambiguous:

Human presence alone does not constitute meaningful oversight where the output cannot be checked at usable cost.

Verification carries a cost. It requires evidence, provenance, diagnostic information, sufficient time, and the authority to challenge the model. Where a reviewer has ninety seconds, no access to the system's reasoning, and a substantial queue of outstanding cases, the oversight step exists on the process map but not in practice. The control has been specified without being resourced, and a control that cannot be performed is not a control.

3. Explanation does not reliably improve verification

The conventional response to automation bias is to use explainable AI: presenting the reviewer with the AI's reasoning, on the expectation that the reviewer will engage with it. The evidence does not support that expectation.



The systematic literature review by Romeo and Conti found that explanations can increase acceptance without improving verification or accuracy. Naiseh and colleagues, in think-aloud studies of clinical decision support, observed users skipping explanations, reading them selectively, and misinterpreting them – producing both over-reliance and under-reliance even where a full explanation was available. Haag's 2025 meta-analysis identified an overall performance benefit associated with explainable decision support, but indicated that the explanations themselves were not the decisive factor.

Herm's empirical work identifies a plausible mechanism: explanations impose cognitive load. Under time pressure, a reviewer presented with a dense rationale does not examine it more carefully but less.

A field study conducted by Lu and colleagues within a live loan-approval service documented the resulting trade-off. Greater explicitness in AI presentation increased both compliance and measured performance, while reducing active deliberation on precisely the complex cases in which human judgment was expected to contribute most.

An explanation interface may produce the appearance of review while diminishing the engagement it was installed to create.



4. Trust calibration is learned, and persists after conditions change

Wischnewski and colleagues surveyed 96 publications on trust calibration and reached a conclusion at odds with much deployment practice: both over-trust and under-trust degrade performance, and reliability information, training, and interface cues assist only where a user's perceived capability corresponds to the system's actual capability.

Lebiere and colleagues demonstrate how that correspondence is established. Operators learn how far to rely on a system from the relationship they experience between what the interface indicates and what subsequently occurs. Reliance is acquired through experience rather than through documentation.

The implication is significant. Where the system changes – through a model update, a shift in population, distributional drift, or adversarial degradation – interface cues may cease to track true reliability. The operator's learned reliance does not necessarily update. It persists in miscalibrated form at precisely the point at which the system has become less trustworthy than the reviewer believes it to be.

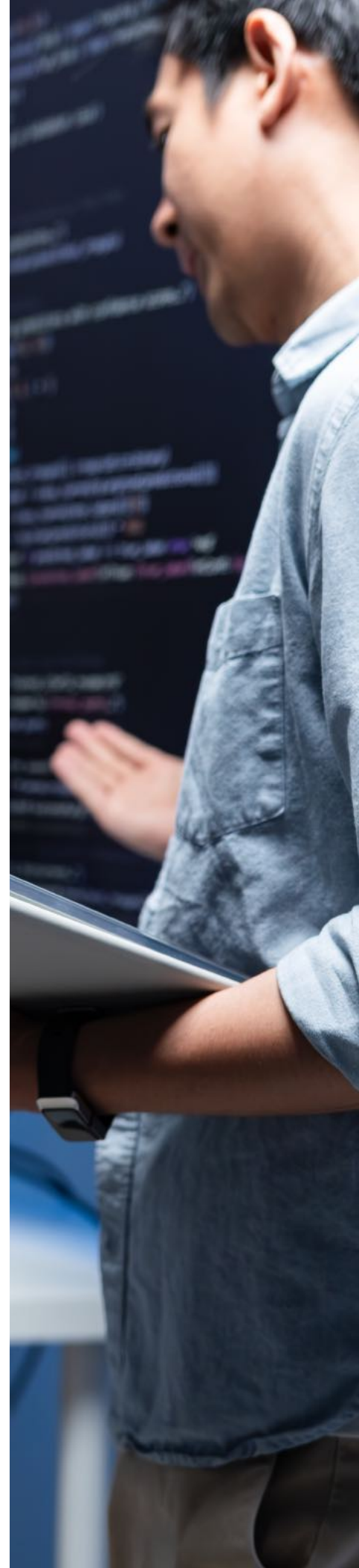
5. Under-reliance is a distinct and equally documented failure

Oversight is not improved by distrust. Systematic reviews by Burton and colleagues (61 articles) and by Mahmud and colleagues (80 empirical studies) document algorithm aversion – the rejection of sound algorithmic advice – as a substantial and separate phenomenon, shaped by expertise, decision autonomy, incentive structures, and the compatibility of the system with established professional reasoning.

Zerilli and colleagues characterize the objective accurately: appropriate reliance occupies the ground between rejection of the system and passive acceptance of it. Vigilance is a narrow band rather than a direction of travel. Neither “trust the system more” nor “trust the system less” constitutes an adequate instruction.

6. Detection and reporting are separate capabilities

Langer, Baum and Schlicker apply signal detection theory to oversight and, in doing so, separate two capabilities ordinarily treated as one: a reviewer's **ability** to detect an inaccurate or unfair output, and their **willingness** to report it. These are distinct failure modes with distinct causes, and a system may fail at either.



A reviewer who cannot perceive the error presents an interface problem. A reviewer who perceives the error and does not raise it presents an organizational problem: authority gradients, production pressure, or a well-founded expectation that raising the matter will prove more costly than permitting it to pass. The two require different remedies, and oversight design typically addresses only the first.

Related evidence points in the same direction. The choice-based conjoint study conducted by Egala and Liang among 362 clinicians found resistance to decision support that constrained the ability to modify an output in accordance with professional responsibility. The meaningful human control framework developed by Cavalcante Siebert and colleagues states the principle directly: a reviewer's responsibility must be commensurate with their practical ability and authority to control the system.

Assigning an individual responsibility for an outcome they lack the time,
information, or authority to influence does not constitute oversight. It
constitutes the pre-positioning of blame.

A Question of Design, Not Diligence

This evidence is readily misread as an account of human deficiency: that reviewers are insufficiently careful, that they ought to examine output more thoroughly, and that the remedy lies in additional training and greater diligence.

That interpretation is both empirically unsupported and practically unhelpful.

Consider the task that a contemporary AI system sets for the person overseeing it. The model is correct in the large majority of instances. It presents its output with uniform confidence irrespective of whether that output is correct. The reviewer is required to sustain vigilance for the rare failure, across a high volume of correct results, under time pressure, and without any signal distinguishing the case requiring attention from those that do not.

This constitutes a request to perform precisely the task that eighty years of research – extending back to Mackworth's radar-operator studies of the 1940s – indicates humans cannot reliably perform. The vigilance decrement is not a deficiency of character. It is a property of human attention, as consistent as any material property an engineer would design around.

A system that depends upon sustained human vigilance to detect rare machine errors is poorly designed. The failure lies in the design, not in the human.

The distinction is not merely a matter of fairness. It separates a tractable problem from an intractable one. Exhortation cannot overcome a cognitive limit. Design can accommodate it – but only where the specific points at which that limit binds are known.

The Conditions for Effective Oversight

The literature converges on a set of conditions. None consists of assigning a human reviewer.

Each of these conditions is measurable. Yet none is currently measured.

Condition	Requirement
Verifiability	The reviewer can check the output at usable cost: evidence, provenance, diagnostic information, and sufficient time in which to employ them.
Calibrated trust	Interface cues correspond to actual reliability, and continue to do so following model updates, drift, or changes in the deployed population.
Authority	The reviewer is able to override the system, and doing so carries no disproportionate professional or procedural cost.
Engagement by design	Consequential cases require active judgment rather than permitting acceptance by default.
Workload headroom	Alert volume, case load, and staffing levels permit the review that the process presumes is taking place.
Psychological safety	A reviewer who identifies a problem is able to raise it without cost – the willingness component of the detection problem.
Failure-mode-specific training	Training addressing how the particular tool fails, in the particular setting, with the particular population – rather than general familiarity with the tool.

The Measurement Gap

The position, as human oversight becomes legally mandatory, may be summarized as follows.

Regulators require the control but collect no data concerning whether it functions. Vendors supply the interface but cannot observe what occurs at the point of use.

Assurance firms audit whether an oversight policy exists – at a point in time, for a single client, under confidentiality – and are not positioned to test whether it works.

Governance platforms instrument the model and are structurally unable to observe the human.

Deploying institutions hold the relevant logs but generally lack both the human factors capability to interpret them and any benchmark against which to do so.

No party can presently answer the most elementary question in AI safety: when a system produces an incorrect output, how often does the human reviewer detect it? No baseline, benchmark, or measurement exists — for any tool, in any setting, in any jurisdiction.

The research literature identifies this gap itself. Mahmud and colleagues, reviewing 80 empirical studies, found the evidence concentrated around technological and individual variables, while task and organizational conditions – workload, escalation norms, incentives, and governance – remained comparatively under-studied. The conditions that most determine whether oversight functions are those least examined.

The gap is nonetheless closable. Festor and colleagues placed intensivists in a high-fidelity simulation, exposed them to unsafe AI recommendations, and employed eye tracking to observe what was in fact attended to and verified. That study establishes the principle: the efficacy of the human safety net can be observed and quantified. It has simply not been undertaken at scale, in the field, or across institutions.

BRITE Institute's Program

BRITE Institute is a nonprofit organization founded on the conviction that trustworthy AI requires infrastructure rather than principles. Human oversight represents the most demanding test of that conviction, and the Institute's human factors and socio-technical research is directed toward it.

Our present or planned activities include:

Measurement of Oversight Efficacy

The Institute's AI Incident Reporting System collects confidential, anonymous reports of near-misses and errors from those closest to AI systems in use, and codes each report for the oversight dimension. Was the error detected or missed? Was it detected by the person formally assigned to oversee the system, or incidentally by another party? Under what conditions: position in shift, alert volume, time pressure, override friction, and whether the system communicated its own uncertainty?

A near-miss report constitutes a record of a human detecting a machine error. An incident report constitutes a record of a human failing to do so. Collected at scale and across institutions, such reports form the measurement instrument that does not presently exist.

Simulation Training and Testing

Simulation training and stress-testing addresses a question no audit answers: whether reviewers will in fact detect AI errors under realistic conditions. Through simulation under controlled pressure, the output is a measured detection rate rather than a policy document. Errors are never introduced into live patient care, and results are never attributed to individuals.



BRITE Institute's program includes collecting data on the effectiveness of current human-in-the-loop systems, analyzing the data to discovering what drives success and failure, and openly publishing such findings.

State of Human Oversight Report

Findings are placed in the public domain through an annual State of Human Oversight report, peer-reviewed research, and open de-identified data. The evidence base upon which oversight is presently designed is thin. The Institute intends to strengthen it publicly, irrespective of whether the findings prove favorable to any party, including the Institute itself.

References

A selection from the research base underpinning this paper. BRITE Institute maintains a fuller review of the human factors literature concerning human-in-the-loop AI.

Burton, J., Stein, M.-K., & Jensen, T. (2019). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*. doi:10.1002/bdm.2155

Carnat, I. (2024). Human, all too human: accounting for automation bias in generative large language models. *International Data Privacy Law*. doi:10.1093/idpl/ipae018

Cavalcante Siebert, L., et al. (2021). Meaningful human control: actionable properties for AI system development. *AI and Ethics*. doi:10.1007/s43681-022-00167-3

Egala, S. B., & Liang, D. (2023). Algorithm aversion to mobile clinical decision support among clinicians: a choice-based conjoint analysis. *European Journal of Information Systems*. doi:10.1080/0960085X.2023.2251927

Faas, C., Kerstan, S., Uth, R., Langer, M., & Feit, A. (2025). Design considerations for human oversight of AI: insights from co-design workshops and work design theory. doi:10.1145/3742413.3789100

Festor, P., Nagendran, M., Gordon, A. C., Faisal, A., & Komorowski, M. (2023). Evaluating the human safety net: observational study of physician responses to unsafe AI recommendations in high-fidelity simulation. doi:10.1101/2023.10.03.23296437

Gaube, S., Langer, M., Miller, T., Baum, K., et al. (2026). Keeping an eye on AI: a framework for effective human oversight of AI systems.

Goddard, K., Roudsari, A., & Wyatt, J. (2011). Decision support and automation bias: methodology and preliminary results of a systematic review. doi:10.3233/978-1-60750-709-3-3

Goulas, S., Megalokonomou, R., & Sotirakopoulos, P. (2026). Why do experts miss AI's errors? Evidence from a randomized labeling experiment. *PNAS Nexus*. doi:10.1093/pnasnexus/pgag146

Haag, F. (2025). The effect of explainable AI-based decision support on human task performance: a meta-analysis. doi:10.48550/arXiv.2504.13858

Herm, L. (2023). Impact of explainable AI on cognitive load: insights from an empirical study. doi:10.48550/arXiv.2304.08861

Hondrich, L., & Ruschemeier, H. (2023). Addressing automation bias through verifiability.

Knop, M., Weber, S., Mueller, M., & Niehaves, B. (2022). Human factors and technological characteristics influencing the interaction with AI-enabled clinical decision support systems. doi:10.2196/preprints.28639

Langer, M., Baum, K., & Schlicker, N. (2024). Effective human oversight of AI-based systems: a signal detection perspective on the detection of inaccurate and unfair outputs. *Minds and Machines*. doi:10.1007/s11023-024-09701-0

Lebiere, C., Blaha, L., Fallon, C., & Jefferson, B. A. (2021). Adaptive cognitive mechanisms to maintain calibrated trust and reliance in automation. *Frontiers in Robotics and AI*. doi:10.3389/frobt.2021.652776

Lu, X., Huang, Y., Zhang, Y., & Shen, L. (2023). Role of presentation explicitness in human-artificial intelligence collaboration: a field study in a loan approval service. doi:10.2139/ssrn.4547893

Mahmud, H., Islam, A., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*. doi:10.1016/j.techfore.2021.121390

Naiseh, M., Cemiloglu, D., Al Thani, D., Jiang, N., & Ali, R. (2021). Explainable recommendations and calibrated trust: two systematic user errors. *Computer*. doi:10.1109/MC.2021.3076131

Romeo, G., & Conti, D. (2025). Exploring automation bias in human-AI collaboration: a review and implications for explainable AI. *AI & Society*. doi:10.1007/s00146-025-02422-7

Schiller, C. A. (2024). The human factor in detecting errors of large language models: a systematic literature review. doi:10.48550/arXiv.2403.09743

Senoner, J., Schallmoser, S., Kratzwald, B., Feuerriegel, S., & Netland, T. (2024). Explainable AI improves task performance in human-AI collaboration. *Scientific Reports*. doi:10.1038/s41598-024-82501-9

Wischnewski, M., Krämer, N., & Müller, E. (2023). Measuring and understanding trust calibrations for automated systems: a survey of the state-of-the-art and future directions. *CHI '23*. doi:10.1145/3544548.3581197

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns*. doi:10.1016/j.patter.2022.100455

European Commission (2024). Regulation (EU) 2024/1689 – Artificial Intelligence Act, Article 14.



Lisanne Bainbridge

By taking away the easy parts of his task, automation can make the difficult parts of the human operator's task more difficult.

