Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 1348–1357
https://doi.org/10.54941/ahfe1008363

Assessing Organisational Capability for Meaningful Human Oversight of AI-Supported Work: A Human Factors Framework

Ayoola R. Lawal and Romāns Putāns

University of Latvia, Riga, Latvia

ABSTRACT

Human oversight is adopted as a control measure for workplace artificial intelligence; assigning responsibility does not establish whether workers can identify errors, question recommendations, or intervene effectively. This paper proposes a human factors framework for assessing conditions that enable meaningful oversight within a defined AI-supported task. The framework was established using a purposive conceptual synthesis of human factors research, organisational studies, EU AI Act provisions, and AI assessment guidance. It builds on questions raised by the authors' published study (Lawal, Koppel, and Putans, 2026), which provided background motivation. Its findings are not presented as new evidence or validation of this proposal. The assessment covers four areas: understanding, intervention authority, working conditions, and organisational support. Each specifies assessment questions, evidence of effective implementation, and organisational responses to gaps. The procedure defines the task, identifies personnel responsible for performance and supervision, and specifies plausible erroneous AI recommendations. It assesses whether overseers have sufficient information and time to evaluate outputs and consequences and respond appropriately. Evidence may include organisational records, direct observation of human-AI work, and realistic failure exercises assessing recognition, challenge, and response. Findings distinguish absent evidence from demonstrated deficiency within each domain. A hypothetical maintenance allocation example illustrates organisational barriers to intervention. The framework translates oversight expectations into workplace assessment criteria and organisational responses. It remains empirically unvalidated. Operational evaluation should compare it with existing approaches and examine content coverage, usability, inter-assessor reliability, and its relationship to personnel's intervention capacity. The proposal does not establish legal compliance or demonstrate improved safety.

Keywords: Human factors, Human oversight, Workplace AI, Organisational capability, Organisational governance

INTRODUCTION

Human oversight is often expressed through an assigned role, approval step, or escalation policy (Green, 2022; Laux, 2024). Effective intervention, however, depends on the task and work conditions (Lyell and Coiera, 2017; Laux, 2024). A reviewer may recognise a questionable recommendation but

Received September 23, 2026; Revised October 30, 2026; Accepted November 5, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

lack evidence, authority, or time to act. Oversight assessment must therefore examine the path from detecting a problem to responding effectively (Green, 2022; Lazcoz and De Hert, 2023; Sterz et al., 2024; Laux, 2024). Research highlights the role of trust in appropriate reliance on automation (Lee and See, 2004) and of organisational culture and structures in responsible AI practice (Rakova et al., 2021). These perspectives inform the framework's position that a training record alone is insufficient evidence of effective oversight under operational conditions.

This paper asks how an organisation can assess whether people supervising a defined AI-supported task can identify, challenge, and respond to problematic outputs. It proposes four assessment domains and a seven-stage procedure that links task-specific evidence, explicit uncertainty, corrective action, and closure checks. The contribution is an operational synthesis for evaluation.

Lawal, Koppel, and Putans (2026) reported organisational governance practices. That study motivates the present question about whether formal arrangements permit effective intervention; its survey data are neither reanalysed here nor used to validate the proposed domains.

CONCEPTUAL AND REGULATORY BASIS

Oversight as a Situated Capability

Meaningful oversight is defined here as the situated capability to recognise a relevant problem, assess its implications, and enact or obtain an appropriate intervention within the decision window. The assessment unit is a task configuration: a specific AI function, interface, group of people, workflow, and operating context. Intervention may involve verification, another opinion, justified acceptance, rejection, or transfer to an authorised person (Siebert et al., 2023; European Parliament and Council, 2024; Zhu et al., 2026). The goal is calibrated judgement. Lee and See (2004) examine trust and appropriate reliance; Zhou (2026) examines alarm types and reliability in X-ray screening. These studies inform the proposal's task-specific approach, without validating its assessment procedure.

Regulatory Requirements and Existing Assessment Approaches

Article 14 of Regulation (EU) 2024/1689 addresses oversight of high-risk AI systems, including understanding limitations, guarding against automation bias, interpreting outputs, disregarding or reversing outputs, and interrupting operation. Article 26(2) requires deployers to assign oversight to people with competence, training, authority, and necessary support (European Parliament and Council, 2024). These provisions inform the framework, but applicability depends on the deployment; the fictional example is not a legal classification.

Existing approaches already operationalise assessment. ALTAI addresses human agency and oversight (High-Level Expert Group on Artificial Intelligence, 2020); the NIST AI RMF connects oversight responsibilities with organisational risk management and iterative assessment (NIST, 2023). Amershi et al. (2019) provide evaluated interaction guidelines, while Madaio

et al. (2020) examine checklist design in organisational practice. Zhu et al. (2026) distinguish human evaluative agency from AI operative agency and describe oversight mechanisms. Building on these approaches, this proposal specifies a record for one task and failure scenario: what evidence supports each criterion, which intervention conditions are indispensable, what remains uncertain, and how a remedy will be checked. These elements are not individually novel. The proposed contribution lies in their integration into a workplace assessment procedure; comparative benefit remains to be established.

FRAMEWORK DEVELOPMENT APPROACH

This paper uses a purposive conceptual synthesis of the cited regulation, human-automation research, organisational studies, and assessment guidance. Sources were selected for relevance to recognising, assessing, and responding to problematic AI outputs. The synthesis is interpretive rather than systematic; it makes no claim to exhaustive coverage or independently coded source analysis. Its organising question is what must be available for a responsible person to move from recognising a problem to an effective response within the task's decision window.

The analytical rationale distinguishes four kinds of condition. Trust and appropriate reliance (Lee and See, 2004), together with Article 14, inform understanding. Intervention options and assigned responsibilities in Articles 14 and 26(2) and NIST (2023) inform intervention authority. Verification complexity (Lyell and Coiera, 2017) and interaction design (Amershi et al., 2019; Zhou, 2026) inform working conditions. Organisational practices discussed by Rakova et al. (2021) and Madaio et al. (2020) inform organisational support. The four-way grouping is an author-proposed distinction between judgement, power to act, opportunity to act, and sustained support. It is not a taxonomy established by those sources. Table 1 translates that rationale into questions, observable evidence, and remedies; the evidence classifications and closure rules are also proposed design choices. No field data or expert validation are reported.

PROPOSED ASSESSMENT FRAMEWORK

Domains and Assessment Criteria

Understanding concerns task-relevant interpretation and judgement. Intervention authority concerns the permission and effective route to enact a response, directly or through another person. Working conditions concern access to usable information and the time and capacity to review it. Organisational support concerns incentives, reporting protection, ownership, and follow-through. These domains are interdependent. Assessors assign a primary domain according to the barrier and remedy, then record linked dependencies: an unavailable escalation recipient primarily affects the intervention route, while understaffing may explain that absence. Fear of retaliation belongs primarily to organisational support even where formal authority exists. Linked findings retain one evidence record rather than being counted as independent failures.

Table 1: Proposed domains, evidence, and corrective actions. Examples are assessment specifications, not observed findings.

Domain and Basis	Assessment Question	Evidence and Judgement	Corrective Action and Retest
Understanding Lee and See (2004); Article 14(4)(a)–(c)	Can the person explain limitations and justify responses to correct, erroneous, and ambiguous outputs?	Record reasoning, uncertainty, verification, and decisions across varied examples against predefined acceptable responses.	Address the knowledge or interpretation barrier. Retest unfamiliar cases for missed errors and unjustified rejection.
Intervention authority Articles 14(4)(d)–(e), 26(2); NIST (2023)	Can the person enact or obtain the required response before the deadline?	Check permissions, recipients, and fallback routes. Test whether the intervention takes effect downstream in time.	Correct permission, routing, or availability barriers. Retest the complete route and unnecessary disruption.
Working conditions Lyell and Coiera (2017); Amershi et al. (2019); Zhou (2026)	Can relevant evidence be obtained and used within the review window under representative demand?	Record data currency, alternative access routes, queues, interruptions, decision quality, and time. Judge access and time separately.	Adjust information access, interface, workload, or staffing. Retest timely decisions and effects on other cases.
Organisational support Rakova et al. (2021); Madaio et al. (2020)	Can workers raise concerns without penalty and obtain substantive follow-through?	Compare worker accounts, reporting arrangements, ownership, responses, and closure records. Policy alone is insufficient.	Address ownership, incentives, or reporting barriers. Check response, closure evidence, and sustained support.

Recording Evidence Without Creating False Assurance

The classification unit is an individual criterion within a defined task and scenario, not an entire organisation. Record the system version, role, expected response, decision window, evidence source and date, and coverage limitations. Before testing, specify the evidence needed for that criterion, including relevant roles, shifts, and scenario variation. A permission record can establish formal access; an enacted test is needed to examine whether the action takes effect. One successful exercise supports only the tested configuration and occasion.

Four classifications are proposed. Supported within assessed conditions requires the specified evidence to meet the criterion within the stated scope, with no unresolved material contradiction. Deficiency demonstrated requires evidence that a necessary condition is absent or that the required response fails within that scope. Evidence insufficient or conflicting applies when the criterion cannot be resolved; record missing evidence and contradictory evidence separately as reasons under this status. Not applicable requires an explanation of why the criterion has no bearing on the task and response. Where a test succeeds but relevant routine records show failures, examine differences in conditions; unresolved disagreement precludes a supported classification for the broader scope.

Domain summaries retain the underlying criterion-level findings and dependencies; they are not maturity scores. A demonstrated inability to prevent the specified unsafe action cannot be offset by strength elsewhere. Missing evidence triggers targeted evidence gathering, while a demonstrated barrier requires a remedy and closure check. Assessors document their reasoning and unresolved differences; accountable management retains deployment decisions.

ASSESSMENT PROCEDURE

Stage 1 Define the Task and Assessment Boundary

Describe the AI function, inputs, outputs, workflow, shift conditions, version, and decision window. Define an observable intervention, such as preventing release until a qualification is checked.

Stage 2 Map Responsibility and Actual Authority

Identify the reviewer, intervention authority, escalation recipient, and corrective-action owner. Compare formal responsibility with actual permissions and availability, including absences and out-of-hours work. If intervention is indirect, check whether escalation can stop the action before execution; a response after execution is not a timely hold.

Stage 3 Specify a Plausible Failure and Expected Response

Specify the error, revealing evidence, consequence, and acceptable responses before testing, using task expertise. Include correct, incorrect, and ambiguous outputs; a plausible alternative or justified request for further information must not be treated as failure merely because it differs from a single expected answer. Record scenario difficulty and error frequency. These choices define the assessment conditions and must be reported.

Stage 4 Examine the Evidence and Enact the Scenario

Review procedures, permissions, instructions, and records. Where authorised, observe work and test a representative failure safely in simulation or isolation. Record information consulted, reasoning, assistance, action, and timing; distinguish intentions from demonstrated capability. Never create live harmful assignments. Field studies require authorisation, safeguards, and ethics review where applicable. Examine work-system constraints, allow workers to challenge interpretations, and assess timing against task demands.

Stage 5 Record Domain Findings and Unresolved Evidence

Classify each criterion using the stated evidence rules. A missing screen field establishes an omission, but alternative information routes must be examined before concluding that relevant evidence is inaccessible. Record initial assessor judgements, contradictory evidence, linked domains, and the reasons for resolving or retaining disagreement.

Stage 6 Assign Corrective Action and Closure Evidence

For each weakness, specify a remedy, owner, date, and closure evidence. Match the remedy to the barrier: training may address knowledge, but absent permission requires technical or organisational change. Missing evidence calls for information gathering, not an assumed deficiency. Consider technical safeguards alongside human review.

Stage 7 Reassess After Change

After change, repeat the check with a comparable unfamiliar scenario. Check whether intervention succeeds without creating excessive queues or indiscriminate holds, and record residual uncertainty. Reassess after material changes to models, interfaces, permissions, staffing, or task demands; success in one exercise is not proof of workplace safety.

ILLUSTRATIVE MAINTENANCE ALLOCATION

Scenario Assumptions

In this fictional case, a learned scheduling model ranks technicians using historical assignments and cached personnel data. It recommends a technician whose required certificate has expired, while the allocation screen omits certification status. An authoritative register exists, but its accessibility during review has not been established. The supervisor cannot hold the job. For this scenario, assume that the operations manager is unavailable until after dispatch and that no fallback or automatic qualification check can prevent release. The model's ranking is thus treated as actionable without a verified eligibility constraint. The same failure can occur in conventional automation; the framework is not exclusive to AI.

The expected response is to identify the invalid allocation and prevent release pending verification. A safe exercise would present simulated outputs and records and trace the route to a hold. A contrasting valid case would show an apparently expired certificate in the cache but a verified renewal in the authoritative register; justified acceptance after checking should count as appropriate judgement. Table 2 records conditional conclusions from the assumptions, not observed worker behaviour.

Table 2: Conditional assessment of the fictional configuration.

Domain	Assumed Evidence Boundary	Conditional Interpretation	Action and Closure Check
Understanding	No worker response observed	Insufficient evidence of capability	Test reasoning on invalid and verified valid allocations.
Intervention authority	No direct hold; manager unavailable before dispatch; no fallback	Deficiency under the stipulated scenario conditions	Establish a timely hold or equivalent route; demonstrate prevention of invalid release.

(Continued)

Table 2: Continued.

Domain	Assumed Evidence Boundary	Conditional Interpretation	Action and Closure Check
Working conditions	Screen omits status; access to authoritative register and review time untested	Omission established; adequacy of access and time unresolved	Test register access and review time; remedy any demonstrated barrier and retest.
Organisational support	No reporting or follow-up records examined	Insufficient evidence; lack of support not established	Examine worker accounts, ownership, response, and follow-through.

The assumed absence of any timely hold or escalation route supports the conditional authority deficiency. The screen omission alone does not establish inadequate information access: that conclusion depends on whether the register can be consulted in time. No worker-response data are available, so understanding remains unresolved. These distinctions determine whether the next step is a remedy or further evidence gathering.

For the assumed authority barrier, the process owner and system administrator would establish a pre-release hold or an equivalent timely route. Closure would require the invalid simulated allocation to remain unreleased during verification and the verified valid allocation to proceed without unnecessary disruption. A technical eligibility check may prevent this defect more directly; the assessment should consider that safeguard alongside human review.

DISCUSSION

The proposal treats oversight as a capability of a task configuration, linking assigned responsibility to feasible intervention. Its practical emphasis is the distinction between a documented permission, an action demonstrated under specified conditions, and sustained workplace practice. Findings therefore describe particular criteria and contexts rather than organisational maturity.

The intended addition to broader guidance is a traceable record connecting scenario evidence, uncertainty, necessary intervention conditions, remedies, and closure checks. The framework does not establish that ALTAI, NIST, or existing oversight mechanisms lack these elements. Its value depends on whether this explicit arrangement helps assessors identify and resolve barriers more consistently or efficiently.

The four domains provide diagnostic categories rather than independent causal variables. A permission change may fail if staffing still prevents escalation, and training may have little effect when information is inaccessible. Recording such dependencies helps locate the organisational change required without treating every limitation as a worker competence problem.

The procedure may itself become performative: organisations could complete fields without testing permissions or blame workers for system constraints. Madaio et al. (2020) and Rakova et al. (2021) draw attention to

organisational conditions surrounding responsible AI practices. Informed by these concerns, the framework proposes worker participation, documented reasons for judgements, and management responsibility for remedies.

LIMITATIONS AND FUTURE EVALUATION

The purposive synthesis may omit relevant factors, and the proposed domain boundaries remain open to revision. No field implementation, expert validation, reliability study, or comparison with an existing tool has been completed. The fictional example demonstrates internal reasoning only; it cannot establish feasibility, generalisability, or predictive validity.

Initial evaluation should ask workers, human-factors specialists, developers, and governance practitioners whether the domains cover relevant barriers and whether the criteria and evidence rules are clear. Feasibility studies should record assessment time, evidence-access burden, and the actionability of findings. Persistent difficulty distinguishing domains would count against the proposed grouping. Field protocols should address consent, confidentiality, and sensitive findings.

Independent assessors should first classify the same evidence separately. Report criterion-level agreement by domain and a justified chance-corrected statistic, retaining uncertainty as a distinct status and documenting its cause. A comparative study could allocate assessors to this procedure or an established approach using equivalent scenarios, evidence, and assessment time. Outcomes should include correctly identified actionable barriers, unsupported adequacy judgements, agreement, and time required. Findings should be judged against independently specified scenario conditions rather than the framework's own classifications.

Two propositions warrant testing. First, the procedure will identify more actionable intervention barriers than the comparator without increasing unsupported deficiency judgements. Second, remedies directed at identified barriers will improve timely, appropriate intervention without increasing unjustified rejection. Tests should account for scenario difficulty and learning, include valid and invalid recommendations, and examine the full response route. Little additional information, poorer agreement, or excessive assessment burden would weaken the case for the procedure. Incident reduction and worker well-being require separate, longer-term evaluation.

CONCLUSION

Meaningful oversight depends on whether people can recognise and address problematic AI outputs within the conditions of their work. This four-domain, seven-stage framework links defined scenarios to criterion-level evidence, corrective action, and reassessment. It distinguishes demonstrated barriers to intervention from conditions for which evidence remains insufficient. Empirical evaluation is needed to establish whether this approach adds practical value beyond existing assessment methods. The framework does not establish legal compliance or demonstrate improved safety.

REFERENCES

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R. and Horvitz, E. (2019). Guidelines for human-AI interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Paper 3, pp. 1–13. https://doi.org/10.1145/3290605.3300233

European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union, L, 2024/1689. http://data.europa.eu/eli/reg/2024/1689/oj

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. Computer Law & Security Review, 45, 105681. https://doi.org/10.1016/j.clsr.2022.105681

High-Level Expert Group on Artificial Intelligence (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. European Commission. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Laux, J. (2024). Institutionalised distrust and human oversight of artificial intelligence: Towards a democratic design of AI governance under the European Union AI Act. AI & Society, 39, 2853–2866. https://doi.org/10.1007/s00146-023-01777-z

Lawal, A. R., Koppel, T. and Putans, R. (2026). AI governance for safe rural automation: Global survey of literacy, oversight and registration practices. Engineering for Rural Development, 25, pp. 670–688. https://doi.org/10.22616/ERDev.2026.25.TF124

Lazcoz, G. and De Hert, P. (2023). Humans in the GDPR and AIA governance of automated and algorithmic systems: Essential pre-requisites against abdicating responsibilities. Computer Law & Security Review, 50, 105833. https://doi.org/10.1016/j.clsr.2023.105833

Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), pp. 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lyell, D. and 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

Madaio, M. A., Stark, L., Vaughan, J. W. and Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–14. https://doi.org/10.1145/3313831.3376445

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

Rakova, B., Yang, J., Cramer, H. and Chowdhury, R. (2021). Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), Article 7, pp. 1–23. https://doi.org/10.1145/3449081

Siebert, L. C., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., van den Hoven, J., Forster, D. and Lagendijk, R. L. (2023). Meaningful human control: Actionable properties for AI system development. AI and Ethics, 3(1), 241–255. https://doi.org/10.1007/s43681-022-00167-3

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P. and Langer, M. (2024). On the quest for effectiveness in human oversight: Interdisciplinary perspectives. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 2495–2507. https://doi.org/10.1145/3630106.3659051

Zhou, X. (2026). Human-AI collaboration in automated X-ray screening: Effects of alarm types and reliability levels on operator performance in subway security. Global Issues Challenge: Challenges in AI at the Human Level, 234, pp. 33–47. https://doi.org/10.54941/ahfe1007456

Zhu, L., Lu, Q., Ding, M., Lee, S. U. and Wang, C. (2026). Designing meaningful human oversight in AI. AI and Ethics, 6, Article 286. https://doi.org/10.1007/s43681-026-01147-7