Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Preserving Human Oversight Through Bounded Explanation Deferral in Intelligent Systems

Esther Okpere
University of Sunderland
Comfort A. Akinwalere
Teesside University
Francis T. Omigbodun
Omigboduntobi@gmail.com

Loughborough University
Elizabeth I. Igbodor
Northumbria University
Lukman O. Kolawole
Middlesex University London
Daniel O. Olokpo
Edinburgh Napier University

Research Article

Keywords: Human–AI teaming, adaptive transparency, explanation debt, cognitive workload, human oversight

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11176797/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Adaptive interfaces can reduce immediate information demands by postponing explanations, yet postponed material may remain essential for subsequent human oversight. This paper develops a theoretical framework for bounded explanation deferral in human–intelligent systems integration. It distinguishes an invariant decision-critical information layer from deferrable supporting material and treats the latter as a review obligation with an owner and deadline. A deterministic capacity balance connects present information demand, accumulated explanation debt, recovery opportunities and uncertainty about available attention. Analytical reasoning establishes a necessary cumulative feasibility condition and shows why deferral alone cannot resolve persistent demand. A transparent six-window worked example illustrates how scheduled repayment can remove temporary capacity exceedances while clearing deferred material before a dependent decision. Counterexamples demonstrate failure under insufficient recovery capacity, underestimated reading effort and misclassified critical evidence. These are mathematical illustrations, not participant observations or validated performance effects. The contribution is an auditable specification for adapting information timing while preserving contestability and identifying when operational escalation is required. The framework yields falsifiable propositions for later human-subject evaluation without assuming that lower workload necessarily improves safety.

1 Introduction

An operator supervising an artificial intelligence (AI) engineering system may receive more information precisely when they have the least time to interpret it. During an abnormal event, recommendations, confidence indicators, alarms and diagnostic explanations compete with the task itself. Automation can therefore redistribute difficulty rather than remove it, especially when humans remain responsible for exceptional conditions (Bainbridge, 1983). The design problem concerns the relationship between information and action: what must a person understand now, what can wait, and how does the system ensure that postponed understanding is eventually restored?

Established accounts explain why indiscriminately increasing detail is unhelpful. Cognitive load theory identifies limits on processing during demanding tasks (Sweller, 1988), while multiple resource theory distinguishes interference across processing resources (Wickens, 2008). Neither implies that every reduction in visible material is beneficial. Situation awareness requires perceiving relevant elements, understanding their significance and anticipating change (Endsley, 1995). A concise display can consequently reduce reading effort while removing the evidence needed to recognise an incorrect recommendation. Workload, awareness and decision quality must remain separate design objectives.

Automation research also cautions against treating assistance as a single dimension. Parasuraman, Sheridan and Wickens (2000) distinguish stages and levels of automation, and Parasuraman and Riley (1997) describe different forms of inappropriate use. Kaber and Endsley (2004) examine how adaptive allocation relates to performance, awareness and workload. Onnasch et al. (2014) further identify trade-offs between routine performance and performance during automation failure. These studies motivate a

narrower question than whether adaptation works: under what information conditions can adaptation remain compatible with meaningful human oversight?

Recent work strengthens the case for adaptation without resolving every aspect of its timing. Wang et al. (2025) propose user-centred adaptive transparency for supervisory control. Richardson et al. (2025) investigate physiological prediction of trust, workload and situation awareness, reporting promising results in a ten-participant spacecraft-habitat task. Such research supports investigating context-sensitive assistance, but its existence does not establish that an inferred workload state can safely determine which evidence disappears. Sensor measurements, construct estimates and permission to withhold information are distinct steps requiring distinct justification.

Explanations themselves do not guarantee effective collaboration. Miller (2019) places explanation within social and cognitive processes, while Bansal et al. (2021) show that explanations need not improve complementary team performance. Buçinca, Malaya, and Gajos (2021) show that interventions encouraging deliberation can reduce overreliance by introducing productive tension between convenience and scrutiny. Appropriate reliance involves accepting useful advice and resisting erroneous advice, rather than maximising agreement or reported trust (Lee and See, 2004; Schemmer et al., 2023). Deferring adverse evidence would undermine this distinction even if users reported less strain.

The present paper addresses the temporal consequences of selective disclosure. Explanation debt is defined here as the estimated review effort associated with retained supporting information that has been postponed but remains relevant to a future decision. The term is an accounting construct, not a direct measure of missing knowledge or a claim that cognition behaves like a financial liability. Its purpose is to make a hidden design obligation inspectable: each deferral requires an identified recipient, a review opportunity and a latest permissible completion point. Recovery is part of the adaptation decision itself.

The research question is therefore: under which explicit capacity and information constraints can explanation deferral remain recoverable before dependent decisions? The paper contributes a protected information layer, a bounded ledger of deferred review obligations, and a necessary feasibility condition derived from conservation of review demand. It also supplies counterexamples that prevent the model from being interpreted as a universal advantage for adaptive displays. The scope is theoretical human–systems integration; no participant study, physiological measurement or performance improvement is claimed. This positioning is consistent with the broader need to evaluate the combined human–AI system rather than assume that collaboration necessarily outperforms its components (Vaccaro, Almaatouq and Malone, 2024). The intended outcome is a specification that researchers can challenge, and practitioners can inspect before implementing adaptive interfaces in operational settings.

A motivating case is a maintenance supervisor who receives an AI diagnosis during an equipment fault. Historical traces may wait until stabilisation, but evidence that contradicts the proposed isolation action cannot. Before restart, some previously deferrable traces become essential. The status of information therefore depends on the authorised decision, not its original display category.

2 Theoretical framework

2.1 Information obligations and human authority

Human–AI teaming requires coordination between agents with different knowledge and capabilities. Klein et al. (2004) emphasise the requirements for automation to function as a team player, while coactive design makes interdependence central to joint activity (Johnson et al., 2014). Applied to explanations, these perspectives suggest that an interface should reveal what assistance changes in the operator's responsibilities. A system that silently removes evidence changes the conditions of oversight without necessarily changing the operator's formal accountability. Figure 1 therefore places information obligations between context assessment and display adaptation.

The proposed protected layer contains information whose omission could change an immediate decision: the hazard, current system action, material uncertainty, contradictory evidence, available intervention and relevant time limit. These categories are requirements for domain-specific analysis, not a universally sufficient message template. A summary qualifies only if it preserves the meaning needed for the current decision. Detailed material becomes deferrable only after establishing that the pending action does not depend on it. When that classification is uncertain, the material remains protected until a competent human resolves the uncertainty.

A distinction is also needed between postponing presentation and deleting information. The former retains the original content, provenance, creation time and relationship to the recommendation. The latter destroys a review opportunity. The framework permits only retention-based deferral. It does not authorise suppressing contradictory measurements, concealing weak confidence or replacing an uncertain diagnosis with a confident instruction. The operator can request supporting evidence at any time; requesting it exposes any resulting capacity conflict rather than allowing the interface to revoke the request silently.

2.2 Explanation debt and recovery

Cognitive offloading changes how tasks are distributed between internal processing and external resources (Risko and Gilbert, 2016). A deferred explanation similarly moves part of a task into the environment, but storage alone does not establish that the operator has understood it. Situation awareness recovery requires active reassessment following a disruption (Gartenberg et al., 2014). Accordingly, the ledger is reduced only when the relevant review task is completed, or when an authorised assessment establishes that the obligation is obsolete and records why. Displaying a message or registering a click is not sufficient evidence of comprehension.

Each ledger entry identifies the deferred material, the responsible reviewer, its estimated review effort and a deadline tied to a dependent decision. The same document may contain several obligations for different roles. A maintenance operator might need a concise uncertainty statement before isolating equipment, while a supervisor requires the full causal trace before authorising restart. Their obligations

cannot be merged merely because they reference the same evidence. Transfer at a shift change requires an acknowledged handover and a new capacity check; otherwise, a locally cleared queue can conceal an unresolved team obligation.

Review effort is expressed in estimated seconds of focused attention. This creates an interpretable accounting unit but assumes that effort can be added within a specified task and operator context. Reading speed, expertise, accessibility requirements and interruptions can violate that approximation. Estimates should therefore include contextual explanation and reconstruction effort, not just text length. The numerical example deliberately uses constant effort values to expose the accounting logic. It does not establish a conversion between words, physiological workload and understanding. Table 1 separates the obligations that the model preserves from the assumptions it still needs.

Table 1
Design obligations and unresolved assumptions

Element	Required treatment	Limit
Critical evidence	Present before dependent action	Classification needs domain review
Deferred explanation	Retain source, owner and deadline	Storage does not prove understanding
Review completion	Use task-specific completion criteria	Clicks alone are insufficient
Capacity estimate	Include an uncertainty allowance	Not a direct measure of cognition

2.3 Capacity uncertainty and visible adaptation

Physiological evidence may inform capacity estimates, but it cannot supply an unqualified permission to remove information. Task-evoked pupil responses relate to processing demands (Beatty, 1982), yet pupil size also reflects other influences (Mathôt, 2018). Preprocessing and baseline choices matter (Mathôt et al., 2018). Heart rate variability requires careful interpretation and measurement design (Laborde, Mosley and Thayer, 2017); the low-frequency to high-frequency ratio is not a simple measure of sympathovagal balance (Billman, 2013). These limitations motivate an explicit uncertainty allowance rather than a universal workload threshold.

The display should identify its current presentation mode, show that supporting material is deferred, and permit inspection of the next review deadline. These requirements address mode-awareness problems described by Sarter and Woods (1995) and follow the mixed-initiative emphasis on coordinating assistance with user control (Horvitz, 1999). Amershi et al. (2019) likewise motivate designs that support correction and changing user needs. Adaptation becomes a negotiable change in presentation, with an observable reason and consequence. This does not promise that operators will notice or correctly interpret every transition; transition costs remain part of the capacity calculation.

The distinction also clarifies what a completed obligation means. For a routine trace, checking the relevant event sequence may suffice; for a contested recommendation, the reviewer may need to explain why they rejected an alternative. Completion criteria belong to the task specification and should be set before adaptation begins. Otherwise, the system can appear to repay debt simply by weakening its definition of review. This creates a traceable connection between interface design and the organisation’s existing authorisation procedures.

3 Methodology

This deductive design study combines a focused narrative synthesis with algebraic analysis and a fully specified worked example. Literature was selected for its relevance to automation, workload, explanation, reliance and recovery; the synthesis is not a systematic review. No participant dataset was used; numerical examples represent explicitly stated assumptions rather than empirical observations. The analytical procedure defines obligations, derives a demand balance, examines boundary cases and compares three presentation rules using identical assumed inputs. No experiment, fitted model or computational simulation forms part of the research method.

Let t index fixed decision windows. Ĉ denotes estimated available attention time and ε a non-negative uncertainty allowance, both in seconds. The conservative planning budget is:

$$C_t^- = max(0, \widehat{C}_t - \epsilon_t)$$

1

Mandatory review demand is m; newly arriving deferrable demand is n. The chosen deferral is q, completed review of deferred material is r, and presentation-switching effort is s. All are non-negative attention-time equivalents. D is outstanding review demand at the start of the window. Conservation and bounds give:

$$D_{t+1} = D_t + q_t - r_t; \quad 0 \le q_t \le n_t; \quad 0 \le r_t \le D_t + q_t$$

2

The current presentation is capacity-feasible only if:

$$m_t + n_t - q_t + r_t + s_t \le C_t^-; \quad D_{t+1} \le D_{max}$$

3

Dmax is a design bound on unfinished review, not a validated cognitive threshold. Every entry must also be reviewed by its deadline; aggregate feasibility cannot substitute for this item-level constraint.

Summing equations (2) and (3) across T windows produces the necessary condition:

$$\sum_{t=1}^{T} (m_t + n_t + s_t) \leq \sum_{t=1}^{T} C_t^{-} + D_{T+1} - D_1$$

4

For complete repayment, the final debt D at T + 1 and the initial debt D_1 both equal zero in Eq. (4). This inequality is necessary, not sufficient: capacity available before an item arrives cannot review it, and deadlines may precede later recovery opportunities. The rule protects m□, defers only eligible n□, and allocates spare capacity to the earliest review deadline. An infeasible protected demand or expiring obligation triggers an explicit request for operational intervention; the response itself must be independently assessed.

For reproducibility, each comparison preserves the same arrival sequence, mandatory content and horizon. Deferred material is assumed to remain valid, and review can be divided across windows without additional reconstruction cost. These favourable assumptions isolate the scheduling mechanism. The failure analysis then relaxes capacity or effort assumptions separately, avoiding a composite score that could conceal opposing effects. A result is classified as feasible only when the per-window budget, debt bound and deadline are all satisfied. No claim about human comprehension is inferred from arithmetic feasibility alone.

The worked example uses six windows, zero switching cost and no uncertainty allowance initially. Capacity is (8, 5, 5, 8, 9, 9), mandatory demand (3, 4, 4, 3, 2, 2), and optional arrivals (3, 4, 4, 2, 1, 1). These illustrative seconds create a temporary burst, not estimates from people. D_1 = 0, Dmax = 6, and burst-related obligations expire at the end of window 5. Comparators are immediate presentation, deferral without repayment, and bounded deferral with repayment. Sensitivity analysis varies one recovery budget and review effort through direct arithmetic. No significance test is appropriate because the outputs are deterministic consequences of assumptions, not sampled observations.

4 Results and discussion

4.1 Conditional feasibility and its limits

The first result is a conservation constraint. Moving q□ seconds out of the current display reduces current demand by exactly the amount added to the ledger, before repayment. Across the full horizon, deferral cannot eliminate review effort unless material is legitimately made obsolete, which the worked example does not assume. If all obligations must be cleared, total demand must fit within total capacity. Persistent overload therefore requires another intervention: less incoming work, more qualified capacity, reduced task scope or a justified change in information requirements. Interface compression alone cannot satisfy an impossible balance.

The second result concerns information protection. Eq. (3) has no feasible solution when mandatory demand plus unavoidable switching effort exceeds available capacity, even if all optional arrivals are deferred. This is an impossibility result within the accounting model, not proof that a particular operation is unsafe. It identifies where the interface specification no longer provides an adequate plan. Endsley (2017) cautions against assumptions that increasing autonomy resolves human–automation difficulties. Here, escalation is a declared boundary of adaptation rather than an unmeasured benefit.

4.2 Worked comparison of presentation rules

Table 2 reports all example inputs and the bounded policy outputs. Immediate presentation requires (6, 8, 8, 5, 3, 3) seconds, exceeding capacity by three seconds in each burst window. Deferral without repayment reduces demands to (6, 5, 5, 5, 3, 3), but leaves six seconds of unfinished review beyond the specified deadline. The bounded policy defers three seconds in windows 2 and 3 and repays three in windows 4 and 5. Its demand is (6, 5, 5, 8, 6, 3), with end-window debt (0, 3, 6, 3, 0, 0).

Figures 2 and 3 show different consequences of the same adaptation. Both deferral policies avoid the example's immediate capacity exceedances; only scheduled repayment clears the obligations. Total scheduled review is 33 seconds under immediate and bounded presentation, whereas deferral without repayment schedules 27 and leaves six outstanding. Immediate presentation does not establish completion when its scheduled demand exceeds capacity. These values show no percentage improvement in human performance. They show that evaluating only the burst interval would favour a policy that quietly transfers effort beyond the observation period. Recovery and subsequent decisions must therefore remain inside the evaluation horizon.

Table 2
Complete worked example for bounded deferral

Window	C	m	n	q	r	End debt	Demand
1	8	3	3	0	0	0	6
2	5	4	4	3	0	3	5
3	5	4	4	3	0	6	5
4	8	3	2	0	3	3	8
5	9	2	1	0	3	0	6
6	9	2	1	0	0	0	3

C is available capacity; m mandatory demand; n new optional demand; q deferral; r repayment. All numeric entries except window number are illustrative seconds. End debt is D at t + 1. Switching cost and uncertainty allowance are zero.

4.3 Sensitivity and counterexamples

The example does not establish that bounded deferral always succeeds. Keeping window 4 capacity at eight leaves three seconds of debt for window 5. Since that window also receives three seconds of fresh demand, its capacity must reach six for complete repayment. Figure 4 varies that capacity from four to nine: the remaining debt is two, one, zero, zero, zero, and zero seconds, respectively. The breakpoint follows directly from the stated inputs. It is not an empirically estimated minimum recovery interval or a transferable human performance threshold.

A second perturbation increases the actual effort of all deferred burst material by 50%, so its review demand becomes nine rather than six seconds. It then violates Dmax before recovery starts, although the combined spare capacity in windows 4 and 5 equals nine. With Dmax explicitly enlarged, repayment could be three then six seconds; without that authorised change, the policy must report infeasibility. Table 3 adds missing critical evidence and mandatory overload as counterexamples. These cases demonstrate distinct failure mechanisms that a favourable aggregate workload score could conceal.

Table 3
Counterexamples and their implications

Perturbation	Consequence	Required response
Window 5 capacity = 5 s	One second overdue at deadline	Replan before dependent decision
Deferred effort rises 50%	Peak debt 9 s exceeds bound 6 s	Reassess plan and debt bound
Mandatory demand > capacity	No optional deferral can restore feasibility	Request operational intervention
Contradictory evidence deferred	Timing may fit; oversight is compromised	Reclassify as protected information

Illustrative boundary cases, not observed operational failures.

A further boundary concerns the timing of capacity. Moving spare attention from window 5 into window 1 leaves the six-window total unchanged, yet that earlier capacity cannot review material that has not arrived. Similarly, extending the nominal horizon cannot rescue an obligation whose dependent decision occurs sooner. These counterexamples explain why Eq. (4) cannot certify feasibility by itself. The ledger must preserve arrival times and deadlines, rather than treating every second of capacity as interchangeable.

4.4 Interpretation for human oversight

The framework extends adaptive transparency by attaching future obligations to present simplification. Wang et al. (2025) address adapting transparency to user needs; the present contribution specifies when deferred supporting material must return and when no feasible return schedule exists. The distinction from a generic notification queue is the combination of decision dependence, a protected information layer and completion criteria. A queued message may wait indefinitely; an explanation obligation cannot

remain unresolved beyond the decision that requires it. This is a proposed design contribution, not an assertion of priority over every previous interface architecture.

Lower demand also does not equate to appropriate reliance. A system could satisfy every equation while presenting a misleading summary as mandatory. Trust depends on more than interface timing, including dispositional, situational and learned influences (Hoff and Bashir, 2015). Recovery after a breach may require changes beyond additional explanation (de Visser, Pak and Shaw, 2018). The ledger consequently supports accountability for disclosure but does not certify factual accuracy, causal validity or justified trust. Content assessment and reliance measures remain necessary companions to timing analysis.

4.5 Testable propositions and practical evaluation

Three propositions follow. First, when sufficient recovery capacity exists, bounded deferral should reduce unresolved obligations at dependent decisions relative to otherwise identical deferral without repayment. Second, any immediate workload benefit may be offset by repayment effort, so full-episode workload need not decrease. Third, as estimation uncertainty increases, a conservative budget should produce more infeasibility declarations, potentially preventing some overload but also generating unnecessary interruptions. Observed behaviour can refute these propositions, and they do not assume the proposed interface wins every comparison.

A subsequent evaluation should compare interfaces with matched information content and deliberately include incorrect recommendations. Outcomes should separate decision errors, timely rejection of wrong advice, unresolved obligations, recovery comprehension and perceived workload. NASA-TLX provides a subjective workload measure (Hart and Staveland, 1988), but cannot verify understanding on its own. Appropriate-reliance measures should distinguish acceptance of correct advice from rejection of incorrect advice (Schemmer et al., 2023). Such work would require a justified sample, a prospective analysis plan, and an ethics review; none has been completed for this paper.

Evaluation should additionally record the effort of accessing retained evidence. If expansion requires searching several screens, the apparent benefit may depend on counting presentation effort while omitting retrieval effort. The same observation applies to interruptions introduced by overdue-item reminders.

4.6 Boundaries and organisational implications

The additive effort assumption is the principal modelling limitation. Reviewing two related items together may cost less than reviewing each separately, while reconstructing context after delay may cost more. The framework requires revised estimates when those effects matter; it cannot justify them away through normalisation. Likewise, a scalar attention budget does not represent the resource structure described by Wickens (2008). The present analysis is best used as an auditable planning approximation for a defined task, with conservative estimates, rather than as a physiological law.

Operational responsibility must also be explicit. A team cannot make obligations disappear by transferring them to an already overloaded colleague, and an override cannot create additional attention capacity. Managers should assess whether a recovery period is genuinely available before approving adaptive disclosure. Retaining provenance and limiting physiological data to a necessary purpose would support review without turning workload estimates into employment-performance scores. The model's practical value lies in exposing these obligations early. It provides a reason to redesign work when recovery is infeasible, rather than reporting a temporarily quieter display as successful integration.

5 Conclusion

This paper develops a theoretical account of explanation deferral as a bounded obligation in human–intelligent systems integration. Its central contribution connects immediate information reduction with the capacity and deadlines needed to restore deferred understanding. A protected layer preserves decision-critical evidence, while a review ledger makes postponement visible and assignable. The derived conservation condition shows that deferral redistributes effort over time; it cannot remedy sustained demand that exceeds available capacity.

The worked example demonstrates one feasible schedule and explicit failure conditions, including insufficient recovery capacity, underestimated review effort and incorrectly classified evidence. These results describe the behaviour of stated assumptions, not measured improvements in safety, workload or human performance. Their value is to make adaptive transparency falsifiable and inspectable before implementation.

Further research should test full-episode effects, comprehension at dependent decisions and appropriate reliance when recommendations are wrong. It should also examine whether additive review effort adequately represents different operators and tasks. Intelligent interfaces should be assessed partly by whether they recognise the limits of their own adaptation: preserving oversight requires an achievable opportunity to review what has been deferred, and a clear response when that opportunity does not exist.

Declarations

Clinical trial number

not applicable.

Author Contribution

E.O. contributed to conceptualization, project administration and writing the original draft. Funding acquisition, Visualization, Data Curation, Investigation, Formal analysisC.A.A. contributed to the literature investigation and manuscript drafting. Project administration, Supervision, Writing - Review &

EditingF.T.O. contributed to methodology development and formal analysis. Funding acquisition, Visualization, Data Curation, Investigation, Formal analysisE.I.I. contributed to validation and critical manuscript revision. Funding acquisition, Project administration, Supervision, Writing - Review & Editing, InvestigationL.O.K. contributed to visualization, including preparation of the figures and tables, and verification of the analytical examples. Project administration, Visualization, Data Curation, Formal analysisD.O.O. contributed to supervision and critical review and editing. All authors reviewed and approved the final manuscript and agreed to be accountable for the work.Project administration, Supervision, Writing - Review & Editing, Investigation

Acknowledgement

None

Data availability.

No participant or operational dataset.

References

1. Amershi S et al (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
2. Bainbridge L (1983) Ironies of automation. Automatica 19(6):775–779. https://doi.org/10.1016/0005-1098(83)90046-8
3. Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, Ribeiro MT, Weld DS (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, pp. 1–16. https://doi.org/10.1145/3411764.3445717
4. Beatty J (1982) Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychol Bull 91(2):276–292. https://doi.org/10.1037/0033-2909.91.2.276
5. Billman GE (2013) The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance. Front Physiol 4., Article 26. https://doi.org/10.3389/fphys.2013.00026
6. Buçinca Z, Malaya MB, Gajos KZ (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, pp. 1–21. https://doi.org/10.1145/3449287
7. de Visser EJ, Pak R, Shaw TH (2018) From automation to autonomy: The importance of trust repair in human–machine interaction. Ergonomics 61(10):1409–1427. https://doi.org/10.1080/00140139.2018.1457725

8. Endsley MR (1995) Toward a theory of situation awareness in dynamic systems. Hum Factors 37(1):32–64. https://doi.org/10.1518/001872095779049543
9. Endsley MR (2017) From here to autonomy: Lessons learned from human–automation research. Hum Factors 59(1):5–27. https://doi.org/10.1177/0018720816681350
10. Gartenberg D, Breslow L, McCurry JM, Trafton JG (2014) Situation awareness recovery. Hum Factors 56(4):710–727. https://doi.org/10.1177/0018720813506223
11. Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Hancock PA, Meshkati N (eds) Human Mental Workload. North-Holland, Amsterdam, pp 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9
12. Hoff KA, Bashir M (2015) Trust in automation: Integrating empirical evidence on factors that influence trust. Hum Factors 57(3):407–434. https://doi.org/10.1177/0018720814547570
13. Horvitz E (1999) 'Principles of mixed-initiative user interfaces', Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 159–166. https://doi.org/10.1145/302979.303030
14. Johnson M, Bradshaw JM, Feltovich PJ, Jonker CM, van Riemsdijk MB, Sierhuis M (2014) Coactive design: Designing support for interdependence in joint activity. J Human-Robot Interact 3(1):43–69. https://doi.org/10.5898/JHRI.3.1.Johnson
15. Kaber DB, Endsley MR (2004) The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. Theoretical Issues Ergon Sci 5(2):113–153. https://doi.org/10.1080/1463922021000054335
16. Klein G, Woods DD, Bradshaw JM, Hoffman RR, Feltovich PJ (2004) Ten challenges for making automation a team player in joint human–agent activity. IEEE Intell Syst 19(6):91–95. https://doi.org/10.1109/MIS.2004.74
17. Laborde S, Mosley E, Thayer JF (2017) Heart rate variability and cardiac vagal tone in psychophysiological research: Recommendations for experiment planning, data analysis, and data reporting. Front Psychol 8., Article 213. https://doi.org/10.3389/fpsyg.2017.00213
18. Lee JD, See KA (2004) Trust in automation: Designing for appropriate reliance. Hum Factors 46(1):50–80. https://doi.org/10.1518/hfes.46.1.50_30392
19. Mathôt S (2018) Pupillometry: Psychology, physiology, and function. J Cognition 1(1). Article 16. https://doi.org/10.5334/joc.18
20. Mathôt S, Fabius J, Van Heusden E, Van der Stigchel S (2018) Behav Res Methods 50:94–106. https://doi.org/10.3758/s13428-017-1007-2. 'Safe and sensible preprocessing and baseline correction of pupil-size data'
21. Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. Artif Intell 267:1–38. https://doi.org/10.1016/j.artint.2018.07.007
22. Onnasch L, Wickens CD, Li H, Manzey D (2014) Human performance consequences of stages and levels of automation: An integrated meta-analysis. Hum Factors 56(3):476–488. https://doi.org/10.1177/0018720813501549

23. Parasuraman R, Riley V (1997) Humans and automation: Use, misuse, disuse, abuse. Hum Factors 39(2):230–253. https://doi.org/10.1518/001872097778543886
24. Parasuraman R, Sheridan TB, Wickens CD (2000) A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern Part A Syst Hum 30(3):286–297. https://doi.org/10.1109/3468.844354
25. Richardson EE, Kintz JR, Buchner SL, Clark TK, Hayman AP (2025) Operator-agnostic and real-time usable psychophysiological models of trust, workload, and situation awareness. Front Comput Sci 7:1549399. https://doi.org/10.3389/fcomp.2025.1549399
26. Risko EF, Gilbert SJ (2016) Cognitive offloading. Trends Cogn Sci 20(9):676–688. https://doi.org/10.1016/j.tics.2016.07.002
27. Sarter NB, Woods DD (1995) How in the world did we ever get into that mode? Mode error and awareness in supervisory control. Hum Factors 37(1):5–19. https://doi.org/10.1518/001872095779049516
28. Schemmer M, Kühl N, Benz C, Bartos A, Satzger G (2023) 'Appropriate reliance on AI advice: Conceptualisation and the effect of explanations', Proceedings of the 28th International Conference on Intelligent User Interfaces, pp. 410–422. https://doi.org/10.1145/3581641.3584066
29. Sweller J (1988) Cognitive load during problem solving: Effects on learning. Cogn Sci 12(2):257–285. https://doi.org/10.1207/s15516709cog1202_4
30. Vaccaro M, Almaatouq A, Malone T (2024) When combinations of humans and AI are useful: A systematic review and meta-analysis. Nat Hum Behav 8:2293–2303. https://doi.org/10.1038/s41562-024-02024-1
31. Wang K, Hou W, Hong L, Guo J (2025) Smart transparency: A user-centered approach to improving human–machine interaction in high-risk supervisory control tasks. Electronics 14(3). Article 420. https://doi.org/10.3390/electronics14030420
32. Wickens CD (2008) Multiple resources and mental workload. Hum Factors 50(3):449–455. https://doi.org/10.1518/001872008X288394

Figures

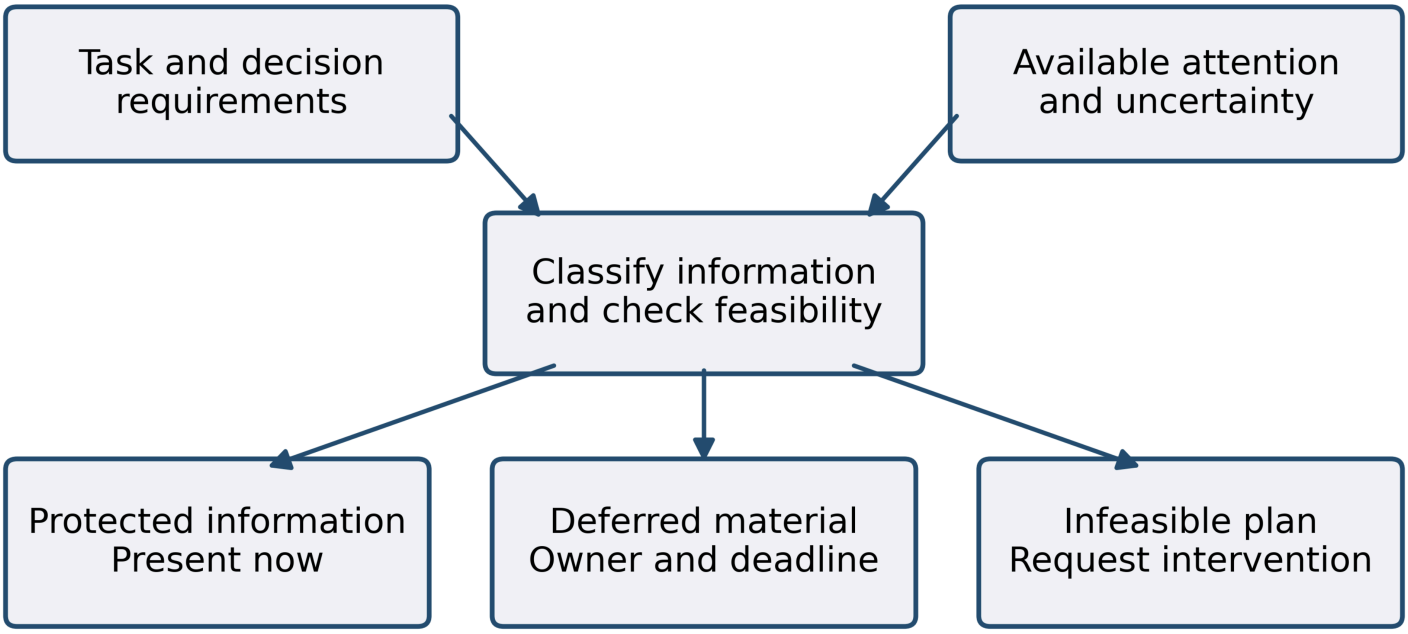


Figure 1

Information obligations govern presentation and recovery. Conceptual design; the protected layer is never eligible for deferral.

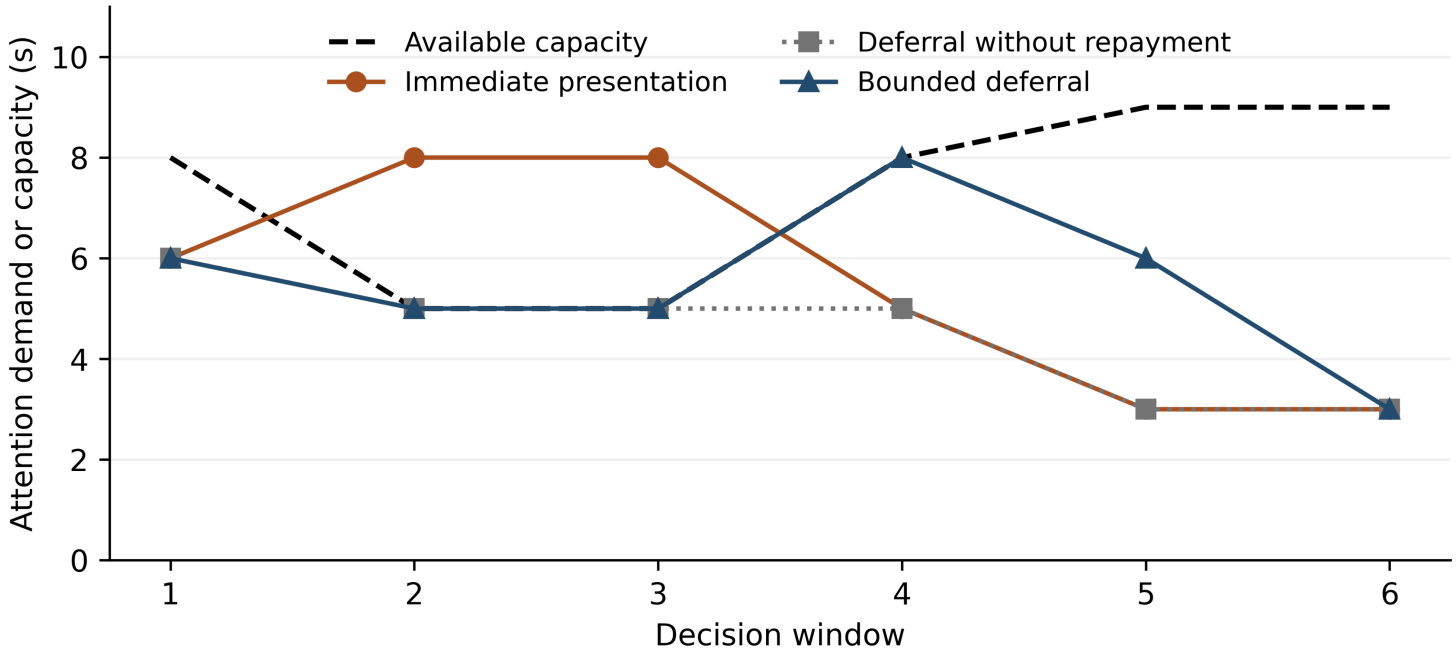


Figure 2

Per-window demand under three presentation rules. All values are illustrative attention-time equivalents from Table 2, not measured workload.

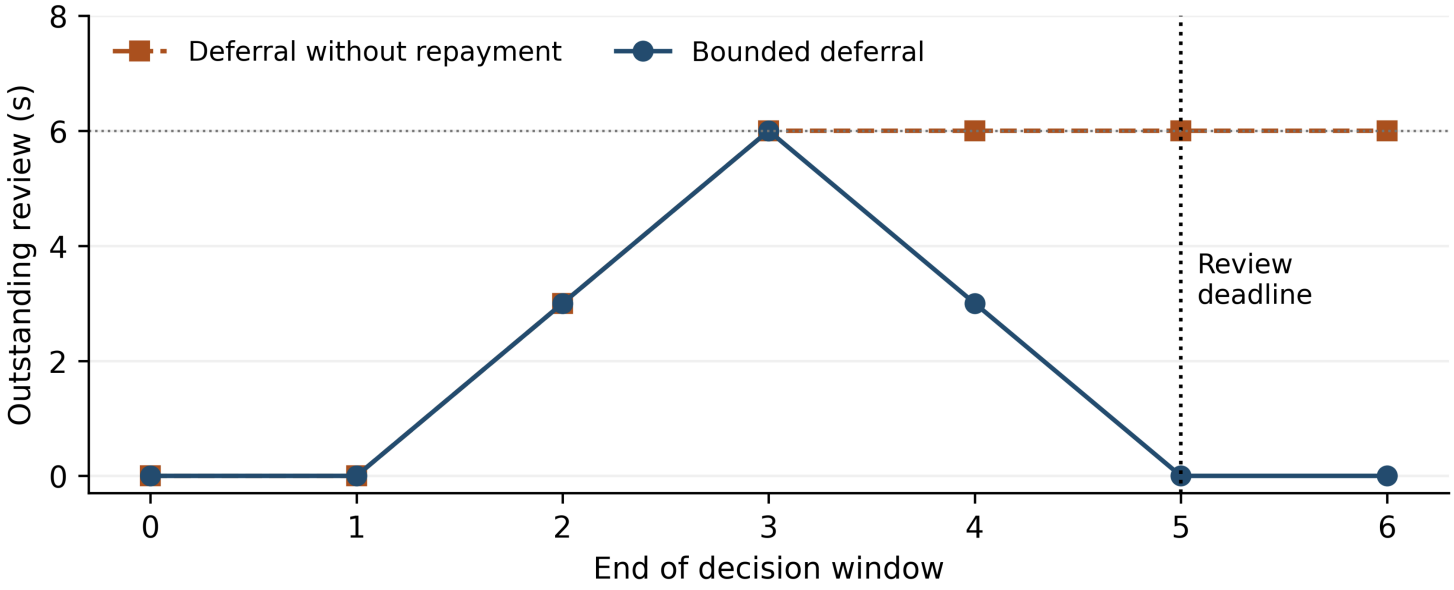


Figure 3

Outstanding review after each window. Both policies defer six seconds; only bounded deferral clears the ledger by the end of window 5—illustrative calculation.

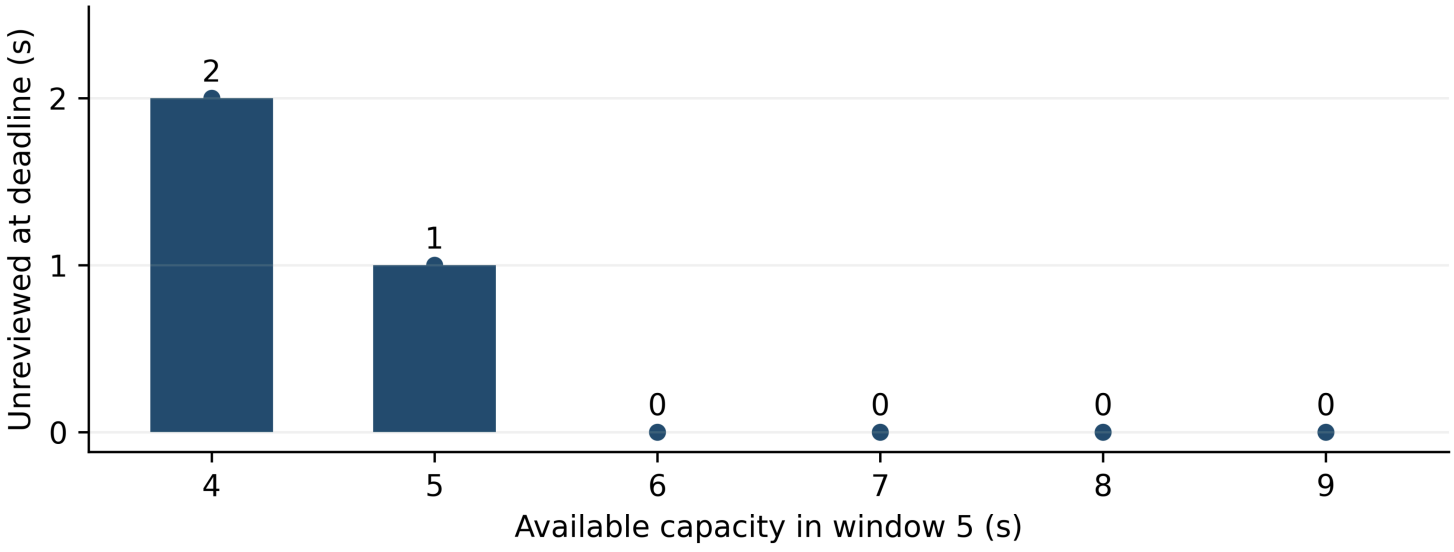


Figure 4

Recovery-capacity sensitivity with three seconds carried into window 5 and three seconds of fresh demand. Values follow max(0, 6 − capacity). Illustrative calculation; other inputs remain fixed.