

Automate the Locks, Gate the Landscapes: Problem-Structure Routing for Human-AI Agent Workflows

Dr Tatia Codreanu

Imperial College London, London SW7 2AZ, UK; I-X Imperial AI in Science Affiliate, London, UK;

Information Cultures, Media Pedagogies, and Screen-Based Research Group, France

Corresponding author: t.codreanu@imperial.ac.uk

Codreanu, T (2026). Automate the Locks, Gate the Landscapes: Problem-structure Routing for Human-AI Agent Workflows (September 17, 2026). Available at SSRN: <https://ssrn.com/abstract=7481979> or <http://dx.doi.org/10.2139/ssrn.7481979>

Abstract

Agentic artificial intelligence shifts automation from generating recommendations to planning and executing multi-step work. This increases the importance of deciding which decisions should be eligible for autonomous convergence in the first place. Existing human-in-the-loop approaches commonly trigger intervention through uncertainty, confidence, risk, or detected failure. This paper argues that these signals are necessary but insufficient because they do not distinguish uncertainty about an answer from legitimate plurality among answers. We introduce Problem-Structure Routing (PSR), a design framework that distinguishes lock problems, whose candidate outcomes can be independently verified against defined criteria, from landscape problems, for which multiple defensible outcomes remain because context, values, stakeholders, or legitimacy affect the choice. Compound workflows contain both and should be decomposed accordingly. The framework is derived through a design-science implementation case involving an AI Storytelling Agent¹. Iterative use showed that explicit brief requirements and output constraints could be treated as verifiable components, while narrative framing, audience fit, tone, and other choices among plausible story alternatives resisted the same automation logic. Corrections therefore separated into execution failures that could be addressed by stronger verification and delegation failures that revealed inappropriate automation of judgment. This distinction led to a workflow redesign in which lock components remained candidates for automation while landscape components received planned human decision gates. The paper formalizes the routing logic, distinguishes execution failure from delegation failure, and proposes an architecture in which agent confidence governs escalation within lock problems but cannot remove human convergence requirements for landscape problems. The result is a practical and testable approach for agent engineering, organizational workflow design, and human-AI governance.

Keywords: agentic AI; human-AI cooperation; automation; human oversight; workflow design; design science; problem structure

1. Introduction

Artificial intelligence is moving from assistance toward action. Agentic systems can decompose goals, plan sequences of steps, call external tools, retrieve and transform information, and execute actions with decreasing levels of direct human control. This shift revives a long-standing question in automation research: which functions should be automated, and to what extent (Parasuraman, Sheridan, & Wickens, 2000)? In agentic systems, however, the question becomes more difficult because a single workflow can combine factual checks, open-ended interpretation, value trade-offs, exceptions, and irreversible commitments.

Human-in-the-loop (HITL) design is often proposed as the solution. However, the phrase remains underspecified. It does not identify which human should intervene, at what point, with what information, or with what authority. Recent research has moved toward more explicit models of meaningful oversight, layered agency, adaptive escalation, and mixed-initiative interaction (Horvitz, 1999; Zhu et al., 2026). Field evidence also shows that intervention quality depends on the type of failure (Wang et al., 2026). These advances address an important operational problem: when should an autonomous system stop or ask for help?

This paper addresses a prior question: which decisions should be eligible for autonomous closure? The central claim is that human intervention should be determined partly by the epistemic and social structure of the problem in addition to model confidence, estimated risk, or observed failure. An agent may be highly confident and technically competent while still operating in a decision space in which several outcomes remain legitimately defensible. In such cases, greater confidence does not create a uniquely correct answer.

To make this distinction operational, the paper introduces two types of problems. A lock problem is a problem for which a candidate solution can be checked against sufficiently explicit criteria without requiring a value judgment about which of several legitimate outcomes should prevail. A landscape problem is a problem for which several outcomes may remain defensible because context, stakeholder interests, values, institutional legitimacy,

¹ <https://asi1.ai/ai/storytellingagent>

or professional judgment affect the choice. Most organizational workflows are compound: they contain lock components embedded within broader landscapes.

The distinction extends earlier work on well-structured and ill-structured problems (Simon, 1973; Reed, 2016) and on tame and wicked problems (Rittel & Webber, 1973), but it serves a narrower implementation purpose. Problem structure is used here as a routing variable for allocating computational convergence and human judgment. The resulting design rule is concise: automate the locks, gate the landscapes, and decompose compound workflows before assigning agency.

The paper defines lock, landscape, and compound problems as an operational task-allocation framework for agentic AI. It also reports a design-science implementation case using an AI Storytelling Agent and shows how repeated correction episodes revealed a difference between failures of execution and failures of delegation. We formalize Problem-Structure Routing (PSR), separating confidence-based escalation from problem-type routing and we derive implications and testable propositions for computer science, business process design, and responsible AI implementation (Codreanu, 2026).

2. Background and related work

2.1 Levels of automation and mixed initiative

Automation research has long rejected the assumption that more automation is always better. Parasuraman and Riley (1997) distinguished appropriate use from misuse, disuse, and abuse of automation. Parasuraman et al. (2000) subsequently proposed that automation can operate at different levels across information acquisition, analysis, decision selection, and action implementation. These arguments can be also applied to agentic AI as automation changes human activity, and the allocation of functions affects monitoring, workload, coordination, and performance.

Mixed-initiative research similarly treats control as a negotiated property of a human-machine system. Horvitz (1999) argued for interfaces in which automated services and direct human manipulation are coupled and are not necessarily treated as mutually exclusive. Human-centered AI research emphasized that high levels of computer automation can coexist with high levels of meaningful human control (Shneiderman, 2020). The design question is therefore linked to the architecture through which initiative and authority are distributed.

2.2 Human oversight in AI and agentic systems

Human-AI interaction research has established the need for systems to make uncertainty, correction, feedback, and control visible (Amershi et al., 2019). Trust research likewise emphasizes appropriate reliance rather than maximal trust (Lee & See, 2004). Users can over-rely on fluent recommendations even when explanations are available. Bućinca, Malaya, and Gajos (2021) found that cognitive forcing functions can reduce overreliance, although interventions can impose usability costs. Bansal et al. (2021) showed that explanations may increase acceptance of AI recommendations regardless of correctness, illustrating why interpretability alone does not guarantee complementary team performance.

With agentic AI, models can initiate and execute actions. Zhu et al. (2026) propose a layered account in which AI has operative agency while humans retain evaluative agency through verification, steering, and substitution. Wang et al. (2026), in a field experiment involving agentic AI in customer service, show that human intervention effectiveness varies with failure type and timing. Mitchell, Ghosh, and Passi (2026) further argue that agent design can erode the cognitive capacities required for effective oversight if humans become passive monitors. Previous research hence shows that human-in-the-loop is neither a sufficient design specification nor a guarantee of meaningful control.

Intervention architectures emphasize variables such as model uncertainty, risk thresholds, anomaly detection, policy violation, or consequences. Those variables help determine whether the system can safely proceed. They do not by themselves determine whether the underlying decision has a single correct resolution.

2.3 Problem structure: from well-structured to plural problems

Simon (1973) distinguished well-structured from ill-structured problems, noting that the boundary is fluid. Well-structured problems have sufficiently specifiable states, operations, and criteria for evaluating solutions. Ill-structured problems leave aspects of the problem space, goals, or evaluation criteria incompletely specified. Reed (2016) revisited the distinction and emphasized that different problem structures demand different combinations of representation, search, domain knowledge, and regulation of cognition.

Rittel and Webber (1973) made a related distinction between tame and wicked problems. For wicked social problems, there is no uncontested stopping rule, and solutions cannot be judged simply as true or false because stakeholders disagree about problem definitions and desirable outcomes. Our lock/landscape distinction builds on this framework. We define a landscape problem as a richly specified problem supported by abundant data, containing legitimate plurality. For example, all relevant evidence about two strategic options may be available while reasonable decision-makers continue to disagree because they weight resilience, fairness, cost, growth, or reputation differently.

In terms of agentic AI, uncertainty and plurality are not the same. Uncertainty concerns whether a system has enough evidence or capability to identify an answer. Plurality concerns whether one answer should be treated as uniquely correct. An uncertainty estimator can reduce the first problem. It cannot eliminate the second without importing a value function, stakeholder priority, or institutional authority that must itself be justified.

3. Method: design-science implementation case

The study follows a design-science orientation in which knowledge is generated through the construction, use, and iterative refinement of an artifact (Hevner, March, Park, & Ram, 2004; Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007). We do not report a controlled benchmark of model performance. Our aim is to derive and specify a reusable workflow design principle from an implementation problem encountered while configuring an AI agent for delegated knowledge work.

The implementation setting was an AI Storytelling Agent configured to support a structured storytelling workflow. The agent could receive a brief, generate or transform narrative material, follow explicit output instructions, and iterate in response to user direction. In storytelling, the same workflow combined components with clear pass-fail criteria and components for which quality depended on context and interpretation. Required brief elements, requested structure, length, and explicit inclusion or exclusion instructions could be checked. Narrative angle, emphasis, tone, audience fit, and the relative weight given to different plausible story directions could not be validated in the same way without first making a judgment about which legitimate alternative should prevail.

During iterative use, human corrections were analyzed according to what kind of problem they exposed. Some corrections concerned compliance with explicit requirements. These remained appropriate candidates for automation because a validator or reviewer could determine whether the requirement had been met. Other corrections arose when the agent produced a plausible but contextually unsatisfactory framing, narrowed the story too early, or resolved a creative trade-off that remained legitimately contestable. In those cases, improving the prompt could change the answer, but it did not convert the underlying decision into a uniquely verifiable one. This difference became the basis for the lock and landscape distinction.

The design response changed from a generic HITL approach, in which a human reviews agent output after the fact, to a routing architecture in which human intervention is planned before deployment. The unit of analysis also changed from the storytelling workflow to the decision component. Three iterative design moves followed: (1) classify components by verifiability and legitimate plurality; (2) decompose mixed tasks into lock and landscape components; and (3) assign different convergence rights to the agent and the human. Table 1 summarizes the decomposition that motivated the framework.

The evidentiary status of the case is a researcher-led design case based on iterative agent construction, observed outputs, corrections, and workflow redesign. Future research will aim to estimate population effects, compare models statistically, and establish the proposed categories. Current research identifies a recurrent implementation problem, creates an explicit artifact to address it, and yields propositions that can be tested in future controlled and field studies.

The Storytelling Agent is used here only as the implementation substrate from which the routing distinction was derived. This paper does not specify or evaluate the agent's broader trust-oriented architecture, multi-agent decomposition, trust-calibration instrumentation, behavioral metrics, or validation protocol. We aim to publish those elements in a future paper.

Storytelling workflow component	What can be checked	PSR classification	Routing implication
Required brief elements and explicit output constraints	Presence, requested structure, length, inclusion or exclusion rules	Lock	Agent may execute; verify compliance and escalate exceptions.

Narrative generation	No single option is uniquely correct if several coherent directions satisfy the brief	Landscape	Agent widens the option space; human retains choice of direction.
Tone, emphasis, and audience fit	Assessment depends on context, audience, intention, and trade-offs	Landscape	Agent analyses or proposes alternatives; human closes the decision.
Final story production	Combines checkable constraints with interpretive narrative choices	Compound	Decompose. Automate verifiable checks; gate narrative convergence.

Table 1. Problem types and routing implications

4. Lock, landscape, and compound problems

4.1 Definitions

A lock problem has a sufficiently determinate evaluation function. Given the relevant input and criteria, an independent evaluator can determine whether a candidate output satisfies the task without first choosing among competing value systems. Examples include checking whether a document contains required fields, reconciling a calculation, determining whether an explicit threshold is met, extracting specified information from a trusted source, or validating a format against a schema. Lock problems can still be difficult, probabilistic, or high stakes. What defines them is the availability of a legitimate verification procedure.

A landscape problem is characterized by legitimate plurality. Two competent decision-makers may examine substantially the same evidence and still choose different outcomes without either choice being reducible to factual error. The difference may arise from context, weighting of objectives, stakeholder interests, professional judgment, values, or institutional legitimacy. Examples include choosing a strategy, deciding whether an exception is justified, balancing cost against resilience, selecting a candidate for overall fit, or deciding what level of risk an organization should accept.

A compound problem combines both. Procurement, for example, may contain lock components such as verifying insurance coverage, delivery constraints, or mandatory certifications. The decision about which qualifying supplier to select may remain a landscape because strategic fit, resilience, sustainability, reputation, and relationships must be weighed. Treating the entire workflow as either automatable or non-automatable obscures this structure.

4.2 A practical classification test

A practical classification test is: could two competent experts, with access to the relevant evidence, defensibly disagree about the final answer? If the answer is no and a verification procedure exists, the task is likely to contain a lock. If the answer is yes because the disagreement depends on legitimate weighting, context, or values, the task contains a landscape. The aim of this test is pragmatic as it asks what kind of disagreement remains after factual errors and missing information are corrected.

The classification is not necessarily definitive. A landscape can contain lockable subproblems, and some activities can become more lock-like when organizations explicitly codify objectives and acceptable trade-offs. At the same time, apparent locks can reveal hidden landscapes when exceptions, conflicting stakeholders, or ambiguous goals surface. The framework therefore supports reclassification during monitoring.

5. From agent errors to workflow redesign

5.1 Two different meanings of failure

The Storytelling Agent made the distinction between execution error and delegation error visible. An execution error occurred when the agent failed to satisfy an explicit requirement that remained independently checkable. In the storytelling workflow, these included failures against stated brief elements, structural constraints, or other explicit output instructions. The appropriate response was primarily technical: improve instructions, constraints, validators, test coverage, retries, or exception handling. The task itself remained a lock because success could still be verified without choosing among legitimate narrative alternatives.

A delegation error occurred when the workflow implicitly treated a narrative judgment as if it had the same kind of answer. A story could satisfy the explicit brief and still be wrong for the user's purpose because the agent selected a framing, emphasis, tone, or narrative direction that was only one of several defensible

possibilities. The problem was not necessarily low model confidence or factual inaccuracy. It was that the workflow had granted the agent authority to close a choice whose adequacy depended on contextual judgment. Prompt refinement could make the output more persuasive without resolving this allocation problem.

This changed the role of the human in the Storytelling Agent workflow. Human involvement initially appeared mainly as correction after automation. The redesign moved the human decision point upstream. Verifiable requirements remained candidates for autonomous execution and verification, while interpretive story choices were treated as landscape components in which the agent could generate, analyze, compare, or challenge alternatives but could not determine final convergence by default. Post-hoc correction therefore became planned human intervention by design (table 2).

Failure class	Type of failure	Typical symptom	Primary design response
Execution failure	Agent execution within a genuine lock	Wrong extraction, calculation, lookup, format, or rule application	Improve engineering, verification, validators, tools, retrieval, retries, and exception handling.
Boundary failure	A landscape component was treated as a lock	Agent silently resolves ambiguity or competing objectives	Decompose task and insert a human decision gate.
Convergence failure	Agent closes alternatives too early	One polished recommendation suppresses defensible options	Require option preservation, assumptions, trade-offs, and human closure.
Verification failure	Automation cannot be checked efficiently	Human must effectively redo the task to trust the output	Redesign verification or reconsider whether automation creates net value.

Table 2. Failure classes and design responses

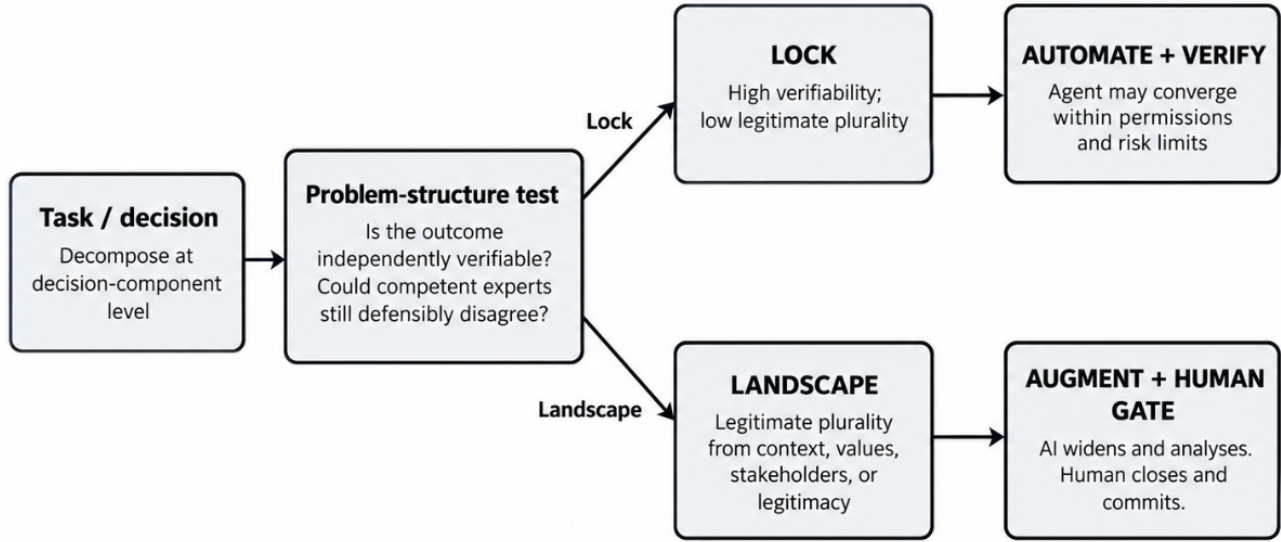
5.2 Confidence is not enough

Confidence-based escalation is valuable inside lock problems. If an agent can estimate that a classification, retrieval, or extraction is unreliable, uncertainty can trigger review. Risk thresholds can similarly require approval before a consequential action. However, confidence and risk do not define whether a landscape should be closed by the agent.

The conceptual difference can be stated directly: uncertainty asks whether the system knows the answer; plurality asks whether there is one answer for the system to know. A system can be highly confident about all relevant facts and still face a normative or contextual choice. Therefore, an architecture that reduces human intervention as model confidence increases can be appropriate for locks but unsafe as a general rule for landscapes.

6. Problem-Structure Routing architecture

Problem-Structure Routing (PSR) treats classification as a control layer that sits before or alongside agent planning. The system first decomposes a workflow into decision components. Each component is assessed on at least three dimensions: verifiability, legitimate plurality, and consequence. The resulting route determines whether the agent may converge, must seek approval, or should remain in an augmentative role (figure 1).



Compound workflows contain both types. Decompose, delegate the locks, gate the landscapes.

Figure 1. Problem-Structure Routing for compound human-AI workflows.

6.1 Formal routing variables

Let t denote a decision component. Define $V(t)$ as the degree to which a candidate output can be independently verified against explicit criteria. Define $P(t)$ as the degree of legitimate plurality, meaning the extent to which competent decision-makers may reasonably select different outcomes after factual discrepancies are resolved. Define $C(t)$ as consequence or irreversibility. These variables are conceptually independent: a task can be highly verifiable and highly consequential, while a landscape can have low factual uncertainty but high plurality.

A simplified routing policy is:

If $P(t)$ is high: AI = augment; human gate = required.

If $P(t)$ is low and $V(t)$ is high: AI = automate/execute, subject to risk and permission thresholds.

If both $P(t)$ and $V(t)$ are unclear: decompose, request missing information, or escalate.

Agent confidence, denoted $c(t)$, is then used within the selected. For example, low $c(t)$ in a lock problem can trigger review, while high $c(t)$ in a landscape problem does not eliminate the human gate. Consequence $C(t)$ can also add approval requirements to a lock problem. A verified calculation may still require human authorization before an irreversible payment or regulatory submission.

6.2 Mandatory human gates

In practice, four classes of events are strong candidates for human gates: commitment, consequence, exception, and values. Commitment concerns actions that bind the organization or user. Consequence concerns actions whose failure creates material harm or difficult reversal. Exception concerns states outside the tested envelope or policy. Values concern choices among competing legitimate objectives. These gates are not all landscape problems in a strict epistemic sense, but they identify points where accountable human authority should remain visible.

The human gate must be operational, the responsible person needs sufficient information to understand the proposed action, the evidence and assumptions behind it, the alternatives that remain, and the consequences of approval. They also need real authority to stop, alter, or redirect execution. Otherwise, the system retains nominal HITL while effectively shifting decision authority to the agent.

7. Implications for computer science and AI implementation

7.1 Separate the planner from the routing policy

Agent architectures commonly focus on planning, tool use, memory, reflection, and evaluation. PSR adds a distinct question: should this component be planned toward autonomous convergence? The routing policy should therefore be separable from the task planner. This separation creates a place to encode organizational constraints, problem-type rules, and human authority without requiring the model to infer them implicitly on every turn.

7.2 Build verifiers for locks

Lock automation depends on efficient verification. Where verification requires a human to repeat the entire task, apparent automation may merely move work from production to checking. For engineering teams, the appropriate investment is therefore often a task-specific verifier: schema validation, deterministic recomputation, source matching, constraint checking, dual execution, or targeted exception tests. The better the verifier, the more safely the human can move from continuous review to exception handling.

7.3 Preserve alternatives for landscapes

Landscape support requires a different evaluation target. The system should be judged on whether it surfaces materially different options, identifies assumptions, separates evidence from inference and value judgments, represents stakeholder conflicts, and preserves contestability until a human decision point. In landscape tasks, premature convergence can be a design failure even when the final recommendation appears plausible.

7.4 Evaluate routing errors explicitly

Traditional evaluations ask whether the agent completed the task correctly. PSR suggests adding routing accuracy. A false-lock error occurs when a landscape is misrouted to autonomous convergence. A false-landscape error occurs when a verifiable task is unnecessarily escalated, creating friction and cost. These errors have asymmetric consequences and can be measured separately from ordinary task accuracy. This creates a research agenda around classifiers, orchestration policies, and benchmark datasets for problem-structure routing.

8. Implications for organizations and business process design

For organizations, the framework changes the unit of AI transformation. Roles and end-to-end processes are often too coarse for automation decisions. A more useful unit is the decision component. Teams can map where evidence is gathered, where criteria are applied, where alternatives are weighed, and where commitments are made. This makes it possible to automate high-volume locks while protecting the judgment points that give the workflow legitimacy.

This also changes the meaning of an automation target. The percentage of a workflow executed by AI is not a sufficient measure of success. A system that automates a certain amount of verifiable work and preserves high-quality human judgment may be preferable to a system that automates a large amount and absorbs contextual decisions into opaque model behavior. The objective is appropriate delegation with measurable verification cost, decision quality, escalation load, and accountability.

Organizations should document at least five properties for each agentic workflow: the tasks eligible for autonomous execution, the verifier used for each lock, the events that force escalation, the human role authorized to close each landscape, and the evidence retained for reconstruction after failure. These properties convert human-in-the-loop from into a process specification.

9. Propositions and research agenda

The framework yields testable propositions and future research will aim to a complete empirical validation.

- P1, Verification proposition. For tasks with high verifiability and low legitimate plurality, task-specific automated verification should permit greater agent autonomy with lower continuous human review than generic HITL, without reducing correctness.
- P2, Plurality proposition. For tasks with high legitimate plurality, increasing model confidence or average task accuracy should not, by itself, justify removal of human convergence authority.

- P3, Decomposition proposition. In compound workflows, component-level routing should reduce both unnecessary human intervention and inappropriate autonomous convergence compared with uniform workflow-level automation policies.
- P4, Failure-type proposition. Distinguishing execution failures from delegation failures should lead to more effective remediation because the former primarily require engineering changes while the latter require changes in allocation of agency.
- P5, Oversight-quality proposition. Human gates will be more effective when they occur before commitment and receive structured evidence, assumptions, alternatives, and consequences.

Future research can test these propositions through controlled task experiments and field deployments. Useful dependent variables include task accuracy, routing error rate, verification effort, human correction time, escalation rate, decision diversity, trust calibration, subjective agency, and downstream consequence severity. Inter-rater reliability should also be measured for lock/landscape classification. If experts cannot classify a task consistently, the appropriate response may itself be to treat the task as a landscape or compound problem until its decision criteria are clarified.

A second research direction concerns dynamic reclassification. As organizations codify policy and build better verifiers, some tasks may move toward the lock region, at the same time, new exceptions or stakeholder conflicts may reveal hidden landscapes. Agent systems could therefore maintain a task-structure model that is updated from incidents and human overrides.

10. Limitations

The framework has several limitations. The design case is single-author and implementation-led. It does not establish causal performance gains or generalize across domains. Lock and landscape are analytical categories, many tasks lie on a continuum, and classification can change as evidence, policy, or organizational goals change. Human judgment is not inherently correct, unbiased, or legitimate. Requiring a human gate transfers convergence authority; it does not guarantee good governance. Human decision processes may require their own checks, diversity, review, and accountability.

Verification itself has a cost; a lock is attractive for automation only when the verification regime is reliable and proportionate. If checking the agent requires equivalent expertise and effort to doing the task, the automation case weakens. The framework does not solve value alignment. It makes value-bearing decisions visible and locates responsibility for closing them. Finally, agentic technology is changing rapidly. The specific technical mechanisms for uncertainty estimation, tool verification, and orchestration will evolve, but the underlying distinction between verifiable correctness and legitimate plurality is not tied to a particular model architecture.

11. Conclusion

The transition from generative AI to agentic AI makes delegation architecture a first-order design problem. Human oversight cannot be specified only as a fallback for low confidence, detected risk, or system failure. Some tasks require human intervention because the system is not reliable enough. Other tasks require human intervention because the decision is not reducible to a uniquely verifiable answer.

Problem-Structure Routing makes this difference explicit. Lock problems are candidates for autonomous convergence when verification, permissions, and risk controls are sufficient. Landscape problems may benefit substantially from AI generation, analysis, simulation, and challenge, but the point of convergence should remain a planned human decision. Compound workflows should be decomposed so that both forms of intelligence are used where they are strongest.

The implementation lesson is therefore the following: correct the agent when it fails a lock and redesign the workflow when it is asked to close a landscape. The goal is a deliberate allocation of computation, judgment, agency, and accountability between humans and machines.

Declarations

Ethics. This paper does not report a new study involving human participants or personal data. The implementation material is a researcher-led design case focused on agent configuration and workflow architecture.

Data availability. No new dataset is reported, the paper analyses the design logic of an implementation case it does not aim to publish interaction logs, participant data, or a benchmark corpus. Future empirical validation will require task corpora, routing annotations, and controlled or field comparisons of workflow architectures.

Conflict of interest. The author declares no conflict of interest.

References

- Amershi, S., Weld, D., Vorvoreanu, M., Fournay, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, Paper 3, 1–13. <https://doi.org/10.1145/3290605.3300233>
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems Association for Computing Machinery, New York, NY, USA, Article 81, 1–16. <https://doi.org/10.1145/3411764.3445717>
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), Article 188, 1–21. <https://doi.org/10.1145/3449287>
- Codreanu, T. (2026). Cooperation after the algorithm: Designing human-AI coexistence beyond the illusion of collaboration. arXiv:2602.19629. <https://doi.org/10.48550/arXiv.2602.19629>
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. <https://doi.org/10.2307/25148625>
- Horvitz, E. (1999). Principles of mixed-initiative user interfaces. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 159–166. <https://doi.org/10.1145/302979.303030>
- Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1), 50–80. <https://doi.org/10.1518/hfes.46.1.50.30392>
- Mitchell, M., Ghosh, A., & Passi, S. (2026). AI agents push humans out of the loop. [arXiv:2608.23642](https://arxiv.org/abs/2608.23642)
- Natarajan, M. (2025). Adaptive Agents for Mixed-Initiative Human-AI Collaborations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(28), 29285–29286. <https://doi.org/10.1609/aaai.v39i28.35220>
- Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). "A model for types and levels of human interaction with automation," in *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, doi: 10.1109/3468.844354
- Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
- Reed, S. K. (2016). The Structure of Ill-Structured (and Well-Structured) Problems Revisited. *Educ Psychol Rev* 28, 691–716. <https://doi.org/10.1007/s10648-015-9343-1>
- Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sci* 4, 155–169. <https://doi.org/10.1007/BF01405730>
- Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495–504. <https://doi.org/10.1080/10447318.2020.1741118>
- Simon, H. A. (1973). The structure of ill structured problems. *Artificial Intelligence*, 4(3–4), 181–201. [https://doi.org/10.1016/0004-3702\(73\)90011-8](https://doi.org/10.1016/0004-3702(73)90011-8)
- Wang, Y., Zhu, C., Feng, T., Lu, L. X., & Jia, B. (2026). Agentic AI and human-in-the-loop interventions: Field experimental evidence from Alibaba's customer service operations. [arXiv:2605.14830](https://arxiv.org/abs/2605.14830)
- Zhu, L., Lu, Q., Ding, M. et al. (2026). Designing meaningful human oversight in AI. *AI Ethics* 6, 286. <https://doi.org/10.1007/s43681-026-01147-7>
- Zimmerman, J., Forlizzi, J., & Evenson, S. (2007). Research through design as a method for interaction design research in HCI. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07). Association for Computing Machinery, New York, NY, USA, 493–502. <https://doi.org/10.1145/1240624.1240704>